Chapter 6
Statistical Methods and Measurement

Jarrett Rosenberg

Abstract Useful ways of measuring software engineering phenomena have to address two challenges: defining realistic and valid metrics that can feasibly be collected under the constraints and time pressures of real-world software development contexts, and determining valid and accurate ways of analysing the resulting data to guide decisions. Too often, the difficulties of addressing the first challenge mean that the second is given little attention. The purpose of this chapter is to present different techniques for the definition and analysis of metrics such as product quality data. Specifically, statistical issues in the definition and application of metrics are presented with reference to software engineering examples.

1. Introduction

Measurement is ubiquitous in software engineering, whether for management, quality assurance, or research purposes. Effectively creating and using measurements is critical to success in these areas, yet there is much confusion and misunderstanding about the best way in which to define, collect, and utilize them. This chapter discusses the purpose of measurement and statistical analysis in software engineering research and development, and the problems researchers and practitioners face in using these methods effectively; rather than a "how-to," it is a "when-to." Section 2 discusses some fundamental issues in measurement and the context of measurement. A number of the issues in this section are discussed in the ISO/IEC 15939 standard, *Information Technology – Software Measurement Process*. Section 3 discusses two basic aspects of creating effective measures: metric definition and metric evaluation. Sections 4 and 5 covers methods for description, comparison, and prediction for simultaneous and successive measurements, respectively, whether categorical or numeric. Section 6 returns to the context of measurement in discussing the important topic of data quality.

F. Shull et al. (eds.), *Guide to Advanced Empirical Software Engineering*.

2. Statistics and Measurement

Measurement is the process of assigning labels (typically numbers) to an attribute of an object or action in such a way that the characteristics of the attribute are mirrored in the characteristics of the labels. The assignment process and the resulting numbers are called a *measurement scale* or *metric*. The reverse process is an interpretive one, and thus if the measurement scale is inappropriate, then the corresponding interpretations of its values will be incorrect. In using the terms "measurement" and "metric", it is usually clear from context whether the process or numerical result is being referred to.

The name "statistics" reflects the origin of the field in the collection of demographic and economic information important to the government of the modern nation state. Such measures as the size of the population, the birth rate, and the annual crop yield became important inputs to decision making. The term *descriptive statistics* applies to such measures, whether simple or complex, that describe some variable quantity of interest. Over the past century and a half, the field of *inferential statistics* has been developed to allow conclusions to be drawn from the comparison of the observed values of descriptive statistics to other real or hypothesized values. These inferential methods require some assumptions in order to work, and much of statistical theory is devoted to making those assumptions as flexible as possible in order to fit real-world situations.

2.1. Statistical Analysis and the Measurement Process

Statistical analysis necessarily assumes some *measurement process* that provides valid and precise measurements of some process of interest, as shown in Fig. 1. The results of the statistical analysis are themselves the prerequisite to a decision-

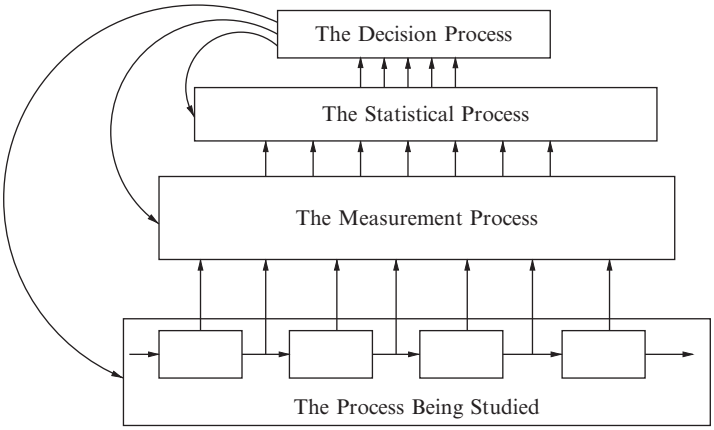

Fig. 1 The roles of the measurement and statistical processes

making process which in turn affects the process of interest, the measurements made on it, and the analyses done on those measurements. It is often the case that too little thought is given to the multi-level nature of this situation: measurements are made because it is possible to do so, statistical analyses are done in a formulaic way, and decisions are made with little data or analysis. In the area of software metrics, Basili et al. (1994) created the "Goal/Question/Metric" framework, which emphasizes that every metric collected must be defined so as to answer some specific question, and every question posed must be relevant to some decision-making goal. This ensures that the entire process depicted in Fig. 1 remains aligned with the overall goal: studying a process in order to make various decisions about it (whether research conclusions or process improvements).

The reason for dwelling on such a banal topic is precisely because it is so often taken for granted; problems with any of these processes or the relations between them become easily lost in the assumption that the overall scheme of things is functioning correctly. Yet if the statistical process is not functioning properly (e.g., incorrect analyses are being performed) decisions will be made on the basis of incorrect analysis and bad outcomes may be misattributed to the decision-making process rather than its statistical inputs. Similarly, it is typically assumed that the measurement process is functioning correctly and that the data it provides are accurate and valid enough to make a statistical analysis worth doing. As Fig. 1 shows, there is no point to a statistical analysis if the data going into it come from a measurement process which is malfunctioning. This involves not only the nature of the measurements involved (discussed in Sect. 3), but also the quality of data obtained.

2.2. The Context of Measurement

While the context of measurement is typically taken for granted and not examined, it nevertheless has a serious impact on the nature and quality of the measurements.

First, the meaning of measurements will vary depending on whether they derive from observation or experiment. If the former, questions of potential bias arise due to various sampling difficulties discussed below. Experiments, on the other hand, while potentially giving precise measurements under controlled conditions, may suffer from a lack of generalizability if they are not carefully designed and interpreted.

Second, it is often the case that the available measurements are not immediately connected with the phenomena of interest: the measures may be what are termed "leading" or "lagging" indicators. The former are highly desirable for forecasting, but the latter are more common; both cases are problematic in steering an organization, because the cause and effect are so separated in time. For example, "number of customer-reported software defects" might seem to be a good metric for evaluating the performance of a software development organization, but it is usually the case that today's customer complaint stems from a defect introduced months or years ago, perhaps by a different set of developers. Similarly, customer satisfaction is typically measured and goaled on an annual or quarterly basis, but it lags a

company's products and services typically by several years. Leading/lagging measures are thus difficult to use in managing day-to-day operations.

Third, while measurements are presumably for a purpose, they can often take on a life of their own, produced because someone once decreed they should be produced, but with no-one paying much attention to them because the rationale has been lost, or is no longer meaningful. Worse, the measurement process can have side-effects, where the numbers are "massaged" or the work process altered in order to produce the "right" results.

Finally, good measurements are *actionable*; they can be used to do something. Measurements made for measurement's sake are worse than useless: they divert resources from the real problems. A single global measure of customer satisfaction or product quality may alert management to a problem, but it gives no indication of what to do. Over time, an organization or researcher will sharpen the questions asked and the corresponding metrics used; this process forms the most important context for measurement and analysis.

3. Creating Effective Metrics

Deciding on an appropriate measure or set of measures is neither as easy as it first appears nor as difficult as it later seems. To be effective, a metric must be clearly defined, have appropriate mathematical properties, and be demonstrably reasonable (i.e., precise, reliable, and valid). Above all, however, a metric must be well-motivated. To be well-motivated, a metric must provide at least a partial answer to a specific question, a question which itself is aimed at some particular research or management goal. For example, how one chooses to measure the time to repair a defect depends on the kind of question being asked, which could range from "What is the expected amount of time for a specific class of defects to go from the initial Reported state to the Repaired state?" to "What percent of all customer-reported defects are in the Repaired state within two days of being first reported?" It is usually the case that a single metric is not sufficient to adequately answer even an apparently simple question; this increases the need to make sure that metrics and questions are closely connected.

3.1. Defining a Metric

Metrics can be either simple or compound in definition. Simple metrics include *counts* (e.g., number of units shipped this year), dimensional *measures* (e.g., this year's support costs, in dollars), *categories* (e.g., problem types), and *rankings* (e.g., problem severity). Compound metrics are defined in terms of two or more metrics, typically combined by some simple arithmetic operation such as division (e.g., defects per thousand lines of code). The number and type of metrics combined and the method

used to combine them affects how easily understood the compound metric will be. This leads to *ratios* (e.g., defects per thousand units), *rates* (time-based ratios such as number of problem reports per month), *proportions or percentages* (e.g., proportion of customers responding "very satisfied" to a survey question), *linear algebraic combinations* (e.g., mean repair cost – the sum of all repair costs divided by the total number of repairs), and *indices* (dimensionless measures typically based on a sum and then standardized to some baseline value). Whereas simple metrics are always defined in terms of some measurement unit, compound metrics such as percentages and some linear combinations and indices can be dimensionless.

The definition of a metric affects its behavior (i.e., the likelihood of its taking on various values), its possible interpretations, and the kinds of analyses which are suitable for it. This argues for the use of simpler, more easily understood metrics rather than the creative development of new, compound ones with poorly understood behavior. Indices in particular raise serious questions of interpretation and comparison, and are best used for showing long-term trends. The range of values a metric can have does not always follow a bell-shaped Normal curve; for example, durations such as repair times almost always have a highly skewed distribution whose tail values pull the mean far from the median. Investigation of the distribution of a metric's values is one of the first tasks that must be undertaken in a statistical analysis. Furthermore, the range of values a measure can take on can be affected by internal or external limitations; these are referred to as truncation or limitation, and censoring.

Truncation or limitation refers to situations where a measure never takes on a particular value or range of values. For example, repair time in theory can never have a value of zero (if it does, the measurement scale is too coarse). Or one may have results from a survey question which asks for some count, with an "*n* or more" response as the highest value; this means that the upper part of the measure is truncated artificially. These situations can sometimes be problematic, and special statistical methods have been developed to handle them (see Long, 1997; Maddala, 1986). A much more difficult case is that of censoring, which occurs with duration data. If the measure of interest is the time until an event happens (e.g., the time until a defect is repaired), then there necessarily will be cases where the event has not yet happened at the time of measurement. These observations are called "censored" because even though we believe the event will eventually occur and a duration will be defined, we do not know how long that duration will be (only that it has some current lower bound). This problem is often not recognized, and when it is, the typical response is to ignore the missing values. This unfortunately causes the subsequent analysis to be biased. Proper analysis of duration data is an extensive sub-area of statistics usually termed "survival analysis" (because of its use in medical research); its methods are essential for analyzing duration data correctly. See Hosmer and Lemeshow (1999) or Kleinbaum (1996) for a good introduction.

Classical measurement theory (Krantz et al., 1971; Ghiselli et al., 1981) defines four basic types of measurement scale, depending on what kinds of mathematical manipulations make sense for the scale's values. (Additional types have been proposed, but they are typically special cases for mathematical completeness.) The four are

Nominal. The scale values are unordered categories, and no mathematical manipulation makes sense.

Ordinal. The scale values are ordered, but the intervals between the values are not necessarily of the same size, so only order-preserving manipulations such as ranking make sense.

Interval. The scale values are ordered and have equal intervals, but there is no zero point, so only sums and differences make sense.

Ratio. The scale values are ordered and have equal intervals with a zero point, so any mathematical manipulation makes sense.

These scale types determine which kinds of analyses are appropriate for a measurement's values. For example, coding nominal categories as numbers (as with serial numbers, say) does not mean that calculating their mean makes any sense. Similarly, measuring the mean of subjective rating scale values (such as defect severity) is not likely to produce meaningful results, since the rating scale's steps are probably not equal in size.

It is important to realize that the definition, interpretation, and resulting analyses of a metric are not necessarily fixed in advance. Given the complexities shown in Fig. 1, the actual characteristics of a metric are often not entirely clear until after considerable analysis has been done with it. For example, the values on an ostensibly ordinal scale may behave as if they were coming from an underlying ratio scale (as has been shown for many psychometric measures, see Cliff, 1992). It is commonly the case that serial numbers are assigned in a chronologically ordered manner, so that they can be treated as an ordinal, rather than nominal, scale. Velleman (1993) reports the case where branch store number correlated inversely with sales volume, as older stores (with smaller store numbers) had greater sales.

There has been much discussion in the software metrics literature about the implications of measurement theory for software metrics (Zuse, 1990; Shepperd and Ince, 1993; Fenton and Pfleeger, 1997). Much of this discussion has been misguided, as Briand et al. (1996) show. Measurement theory was developed by scientists to aid their empirical research; putting the mathematical theory first and the empirical research after is exactly backwards. The prescriptions of measurement theory apply only after we have understood what sort of scale we are working with, and that is often not the case until we have worked with it extensively.

In practical terms, then, one should initially make conservative assumptions about a scale's type, based on similar scales, and only "promote" it to a higher type when there is good reason to do so. Above all, however, one should avoid uncritically applying measurement theory or any other methodology in doing research.

3.2. Evaluating a Metric's Effectiveness

A measure can have impeccable mathematical credentials and still be totally useless. It order for it to be effective, a measure needs an adequate amount of precision, reliability, and validity. One also has to consider its relationships to other

measures, as sometimes misleading results can occur when two related measures are treated as if they were independent.

There are two different concepts sharing the term "measurement precision." One concept is that of the size of a metric's smallest unit (sometimes called its "least count"). Put another way, it is the number of significant digits that can be reported for it. For example, measuring someone's height to the nearest millimeter is absurd, since the typical error in obtaining the measurement would be at least as large. Similarly, measuring someone's height to the nearest meter would be too crude to be of much value. A common mistake is to forget that the precision of any derived measure, including descriptive statistics such as the mean, can not be any greater than that of the original measures, and is almost always less. Thus reporting the average height of a group of people as 178.537 cm implies that the raw measurements were made at the accuracy of 10 μm; this is unlikely. Such a result is better reported as simply 179 cm. The arithmetic combination of measures propagates and magnifies the error inherent in the original values. Thus the sum of two measures has less precision than either alone, and their ratio even less (see Taylor, 1997; Bevington and Robinson, 1992); this should be borne in mind when creating a compound metric.

The other concept of precision is the inverse of variability: the measurements must be consistent across repeated observations in the same circumstances. This property is termed *reliability* in measurement theory. Reliability is usually easy to achieve with physical measurements, but is a major problem in measures with even a small behavioral or subjective component. Rating scales are notorious in this respect, and any research using them needs to report the test-retest reliability of the measures used. Reliability is typically quantified by Cronbach's *coefficient alpha*, which can be viewed as essentially a correlation among repeated measurements; see Ghiselli et al. (1981) for details.

A precise and reliable measure may still be useless for the simple reason that it lacks *validity*, that is, it does not in fact measure what it claims to measure. Validity is a multifaceted concept; while it is conventional to talk about different types of validity, they are all aspects of one underlying concept. (Note that the concepts of internal and external validity apply to *experiments* rather than measurements.)

Content validity is the degree to which the metric reflects the domain it is intended to measure. For example, one would not expect a measure of program complexity to be based on whether the program's identifiers were written in English or French, since that distinction seems unrelated to the domain of programming languages.

Criterion validity is the degree to which a metric reflects the measured object's relationship to some criterion. For example, a complexity metric should assign high values to programs which are known to be highly complex. This idea is sometimes termed *discrimination validity*, i.e., the metric should assign high and low values to objects with high or low degrees of the property in question. In this sense it may be thought of as a kind of "predictive validity."

Construct validity is the degree to which a metric actually measures the conceptual entity of interest. A classical example is the Intelligence Quotient, which attempts

to measure the complex and elusive concept of intelligence by a combination of measures of problem-solving ability. Establishing construct validity can be quite difficult, and is usually done by using a variety of convergent means leading to a preponderance of evidence that the metric most likely is measuring the concept. The simpler and more direct the concept, the easier it is to establish construct validity; we have yet to see a generally agreed-upon metric for program complexity, for example, while number of non-commentary source statements is generally accepted as at least one valid metric for program size.

Finally, a metric's effectiveness can vary depending on its context of use, in particular, how it is used in combination with other metrics. There are three pitfalls here. The first is that one can create several ostensibly different metrics, each of which is precise, reliable, and valid, but which all measure the same construct. This becomes a problem when the user of the metrics doesn't realize that they are redundant. Such redundancy can be extremely useful, since a combination of such metrics is usually more accurate that any one of them alone, but if they are assumed to be measuring independent constructs and are entered into a multivariate statistical analysis, disaster will result, since the measures will be highly correlated rather than independent. Therefore one of the first tasks to perform in using a set of metrics is to ascertain if they are measures of the same or different constructs. This is usually done with a factor analysis or principal component analysis (see Comrey and Lee, 1992).

The second pitfall is that if two metrics' definitions contain some component in common, then simple arithmetic will cause their values to not be independent of each other. For example, comparing a pretest score and a difference score (posttest minus pretest) will yield a biased rather than an adjusted result because the difference score contains the pretest score as a term. Another example is the comparison of a ratio with either its numerator or denominator (say, defect density and code size). Such comparisons may be useful, but they cannot be made with the usual null hypothesis of no relationship (see Sect. 4.2), because they are related arithmetically. This problem in the context of measures defined by ratios is discussed by Chayes (1971), who gives formulas for calculating what the *a priori* correlation will be between such metrics.

The third pitfall is failing to realize that some metrics are not of primary interest themselves, but are necessary covariates used for adjusting the values of other metrics. Such measures are known as *exposure factors* since the greater their value, the greater the likelihood of a high value on another measure. For example, in demographics and epidemiology population size is an exposure factor, since the larger the population, the larger the number of criminals, art museums, disease cases, and good Italian restaurants. Similarly, the larger a source module, the larger the value of any of a number of other metrics such as number of defects, complexity, etc., simply because there will be more opportunity for them to be observed. Exposure variables are used in a multivariate analysis such as Analysis of Covariance (ANCOVA) or multiple regression to adjust for ("partial out") the effect of the exposure and show the true effect of the remaining factors.

3.3. Statistical Analyses

Having defined appropriate metrics and ensured that data is properly collected, the focus shifts to the question of how to appropriately analyze the data obtained. There are three principal statistical tasks involved: *description*, *comparison*, and *prediction*. It is useful to discuss separately the analyses appropriate to dynamic or temporal data, i.e., data which have time as a fundamental aspect, from static data, which do not; however, all statistical analyses have some aspects in common.

The prerequisite for any data analysis is *data cleaning*: the auditing of the data for complete and accurate values. This step typically takes at least as much time, if not more, than the application of the statistical techniques themselves. Often data quality problems prevent many of the intended statistical analyses from being carried out, or create so much uncertainty about the validity of their results as to render them useless. It is usually possible to gather some information from even poor quality data, but an initial investment in data quality pays for itself in the ability to do more – and more useful – analyses later. We will return to this issue in Sect. 6.

Statistical analyses are all based on *models of the underlying data-generating process*; these models can be simple or complex, and can make more or fewer assumptions. *Parametric models* assume specific functional forms such as the Normal distribution for univariate data, or a linear regression equation for multivariate data. The parameters of these functional forms are estimated from the data and used in producing descriptive statistics such as the standard error of the mean, or inferential statistics such as the *t*-statistic used to test for a difference between two means. Because they make stronger assumptions, parametric models can be more useful – if the assumptions are true. If they are not true, biased or even wildly inaccurate results are possible. *Non-parametric models* make few assumptions (typically that the data are unimodal and roughly symmetrical in distribution) and thus can be used in almost any situation. They are also more likely to be accurate at very small sample sizes than parametric methods. The price for this generality is that they are not as efficient as parametric tests when the assumptions for the latter are in fact true, and they are usually not available for multivariate situations.

In the same way that a phenomenon typically cannot be captured by a single metric, a statistical analysis typically cannot be done by conducting one test alone. A good data analyst looks at the data from a variety of different perspectives, with a variety of different methods. From this a picture gradually emerges of what is going on. A word of caution, however: the conventional *p*-value of 0.05 represents a "false positive" or spurious result rate of 1 in 20. This means that the more statistical tests that are performed, the more likely it is that some of them will be falsely significant (a phenomenon sometimes called "capitalization on chance"). Large correlation matrices are a good example of the phenomenon; to see why, compute the 20×20 correlation matrix among 20 samples of 100 uniform random numbers: of the 190 unique correlations, how many are statistically significant at the 0.05 level? It is thus seriously misleading to do dozens of tests and then report a result with a *p*-value of 0.05. The usual way of correcting for doing such a large number

of tests is to lower the *p*-value to a more stringent level such as 0.01 or even 0.001. The most common way of reducing the false positive rate among multiple tests is called the Bonferroni procedure; it and several improvements on it such as the Scheffé and Tukey methods are described in Keppel (1991). Often preferable to multiple univariate tests is a single multivariate analysis.

4. Analyzing Static Measurement Data

4.1. Description

The first step in any statistical analysis is data description, and the first step of data description is to simply *look at the data*. Figure 2 shows the histograms for two different samples with the same mean and standard deviation; without looking at these histograms, one would think from their descriptive statistics that both samples were from the same population. Looking at the distribution of values for a metric allows one to check for most frequent values (modes), outliers, and overall symmetry of the distribution. If a distribution is skewed by a few extreme values (large or small), many widely used statistics become misleading or invalid. For example, the mean and standard deviation are much more sensitive to extreme values than the median or percentiles, and so the mean of a skewed distribution will be far from the median and therefore a somewhat misleading measure of central tendency. Thus looking at the data allows us to determine which descriptive statistics are most appropriate.

As pointed out above, descriptive statistics such as point estimates are subject to error; it is important to quantify this error so that the precision of the point estimate can be determined. The *standard error* of an estimate is a common way of representing

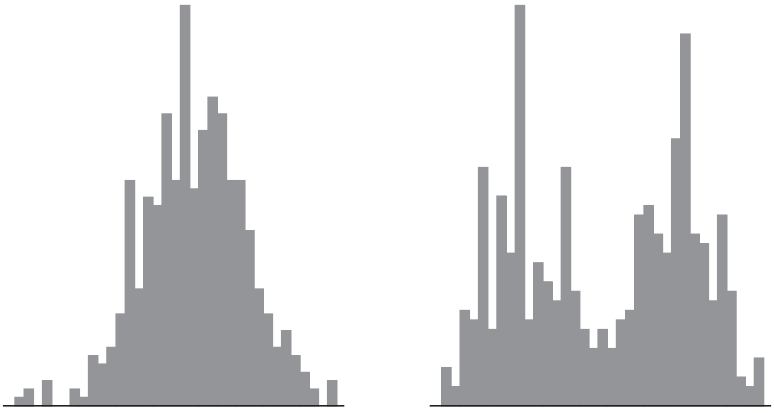

Fig. 2 Two very different samples with the same mean and standard deviation

the precision of an estimate; the range of values two standard errors on either side of the estimate delimit the *95% confidence interval* for that estimate, i.e., the interval within which the true value of the parameter being estimated will fall 95% of the time. A wide confidence interval indicates that the estimate is not very precise, thus knowing the precision is useful for gauging an estimate's value in decision making. The standard error increases as the sample size decreases, and the resulting imprecision in estimates is what makes very small samples so problematic.

4.1.1. Measures of Central Tendency

The main feature of interest in a sample of non-temporal data is its "center of mass". For a roughly symmetric distribution, this will be essentially the same value as its mode (most frequent value) and its median (50th percentile or midpoint). The arithmetic mean is the most commonly used measure of central tendency because of its intuitive definition and mathematical usefulness, but it is seriously affected by extreme values and so is not a good choice for skewed data. The median by definition always lies at the point where half the data are above it and half below, and thus is always an informative measure (indeed, a simple check for skewness in the data is to see how far the mean is from the median). The reason the median is not used more often is that it is more complicated to calculate and much more complicated to devise statistical methods for. When dealing with rates, the geometric mean (the nth root of the product of the n data values) more accurately reflects the average of the observed values.

4.1.2. Measures of Dispersion

Since two entirely different distributions can have the same mean, it is imperative to also include some measure of the data's dispersion in any description of it. The range of the values (the difference between the highest and lowest values) is of little use since it conveys little about the distribution of values in between. The natural measure for distributions characterized by the arithmetic mean is the variance, the sum of the squared deviations about the mean, scaled by the sample size. Since the variance is in squared units, the usual measure reported is its square root, the standard deviation, which is in the same measurement units as the mean. Analogues to the standard deviation when the median rather than the mean is used are the values of the first and third quartiles (i.e., the 25th and 75th percentiles) or the *semi-interquartile range*, which is half the difference between the first and third quartiles. These give a measure of the dispersion that is relatively insensitive to extreme values, just like the median. Another useful measure of dispersion is the *coefficient of variation* (CV), which is simply the standard deviation divided by the mean. This gives some indication of how spread out the values are, adjusted for their overall magnitude. In this sense, the coefficient of variation is a dimensionless statistic which allows direct comparison of the dispersion of samples with different underlying measures (for example, one could

compare the CV for cyclomatic complexity with the CV for module length, even though they are measured in totally different units).

4.1.3. Measures of Association

The most common measure of association between two measures is the correlation coefficient, which is a standardized way of describing the amount by which they covary. The correlation coefficient, *r*, is the square root of the amount of shared covariation between the two measures; thus while r^2 is an easily interpreted ratio measure (an r^2 of 0.4 is half that of an r^2 of 0.8), correlation coefficients are non-linear: an *r* of 0.4 is *not* half that of an *r* of 0.8, but only one-quarter as large. Because they are adjusted for the amount of variation present in the variables being correlated, correlation coefficients among different sets of measures can be compared. However, correlation coefficients are sensitive to the range of variation present in each variable; in particular, large differences in the two ranges of variation place an *a priori* limit on the size of *r*. Thus, special forms of correlation coefficient have been developed for the cases like that of a binary and a continuous variable.

4.1.4. Categorical Data

Categorical data come in two basic kinds: binomial data, where there are only two categories, and multinomial data, where there are more than two. Description of categorical data is typically done by means of the proportion or percentage of the total each category comprises. While pie charts are a common graphical representation, histograms or polar charts (also called Kiviat diagrams or star plots) are more accurately read (Cleveland, 1994). It is important to not report proportions or percentages of small samples to a greater degree of precision than the data warrant: 11 out of 63 cases is not 17.46%, because the smallest percentage that can be observed in a sample of 63 (i.e., one individual) constitutes more than one percent of the sample.

There are a variety of measures of association between two categorical variables (as long as the categories can be considered ordered), see Goodman and Kruskal (1979); all of them can be thought of as special instances of correlation.

4.1.5. Ordinal Data

Ordinal data present special challenges since they contain more information than simple categories, but ostensibly not enough to justify more sophisticated statistical techniques, or even the calculation of the mean and standard deviation. Analysis of ordinal data therefore typically reduces it to the nominal level, or promotes it to the interval or ratio ones. Both of these approaches can frequently be justified on pragmatic grounds.

A prototypical example of ordinal data is the subjective rating scale. The simplest description of such data is simply its distribution, which is done the same way as for multinomial categorical data. Since the number of scale values is limited, simply listing the percentage of cases for each value is more useful than the range or standard deviation. Since such data are often skewed (see Fig. 3 for an example from a satisfaction rating scale), the median is a better measure of central tendency than the mean. Since most responses pile up at one end, this has the effect of making the mean of the scale values most sensitive to changes in values at the other, skewed end (in the case of Fig. 3, at the low-satisfaction end). Thus in Fig. 3 the mean of the satisfaction ratings is paradoxically more sensitive to measuring changes in dissatisfaction than satisfaction.

Correlation of ordinal values is typically done with non-parametric measures such as the Spearman correlation coefficient, Kendall's tau, or the kappa statistic used for inter-rater reliability. Interpretation of such statistics is harder than correlation coefficients because of the lack of equal intervals or ratios in ordinal values; a tau or kappa value of 0.8 is not strictly twice as good as one of 0.4.

4.2. Comparison

Data are rarely collected simply for description; comparison to a real or ideal value is one of the main aims of statistical analysis.

The basic paradigm of statistical comparison is to create a model (the *null hypothesis*) of what we would observe if only chance variation were at play. In the case of comparing two samples, the null hypothesis is that the two samples

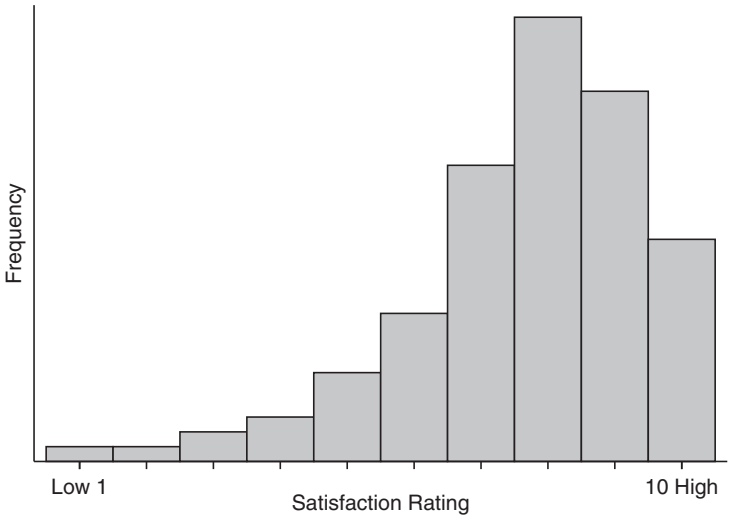

Fig. 3 An example of skewness in ordinal data (from a rating scale)

come from the same underlying population, and thus will have descriptive statistics (e.g., the mean) that differ only by an amount that would be expected by chance, i.e., whose expected difference is zero. If the observed difference is very unlikely to occur just by chance, then we conclude (with some small risk of being wrong) that the two samples are not from the same population, but rather two different ones with different characteristics.

The basic method of statistical comparison is to compare the difference in the average values for two groups with the amount of dispersion in the groups' values. That is, we would judge a difference of 10 units to be more significant if the two groups' values ranged from 30 to 40 than if they ranged from 300 to 400. In the latter case we would easily expect a 10-unit difference to appear in two successive samples drawn from exactly the same population.

Statistical tests of comparison are decisions about whether an observed difference is a real one, and as such, they are subject to two kinds of error:

Type I error (symbolized by α) – incorrectly rejecting the null hypothesis, and deciding that a difference is real when it is not,

Type II error (symbolized by β) – incorrectly not rejecting the null hypothesis, and deciding that a difference is not real when it is.

The probabilities determined for these two types of error affect how a result is to be interpreted. The value for alpha is traditionally set at 0.05; the value for beta is typically not even considered; this is a mistake, because the value of $(1 - \beta)$ determines the *power* of a statistical test, i.e., the probability that it will be able to correctly detect a difference when one is present. The major determinant of statistical power is the size of the sample being analyzed; consequently, an effective use of statistical tests requires determining – before the data are collected – the sample size necessary to provide sufficient power to answer the statistical question being asked. A good introduction to these power analysis/sample size procedures is given in Cohen (1988).

Because of this issue of statistical power, it is a mistake to assume that, if the null hypothesis is not rejected, then it must be accepted, since the sample size may be too small to have detected the true difference. Demonstrating statistical equivalence (that two samples do, in fact, come from the same population) must be done by special methods that often require even more power than testing for a difference. See Wellek (2002) for an introduction to equivalence testing.

The classic test for comparing two samples is the venerable *t*-test; its generalization to simultaneous comparison of more than two samples is the (one-way) analysis of variance (ANOVA), with its *F*-test. Both of these are parametric tests based on asymptotic approximations to Normal distributions. While the two-sample *t*-test is remarkably resistant to violations of its assumptions (e.g., skewed data), the analysis of variance is not as robust. In general, for small samples or skewed data non-parametric tests are much preferred; most univariate parametric tests have non-parametric analogues (here, the Wilcoxon/Mann-Whitney test and the Kruskal-Wallis test). A good reference is Sprent (1993).

Occasionally, one may wish to compare an observed mean against a hypothesized value rather than another group mean; this can be done by means of a one-sample *t*-test or equivalently, if the sample is large (>30), by a *Z*-test.

4.2.1. Categorical Data

Comparison of categorical data between two or more samples is typically done by a chi-squared test on an $n \times m$ table where the rows are the samples and the columns are the categories (see Agresti, 1998; Wickens, 1989). For tables with small cell values (where the standard chi-squared tests are inaccurate), special computationally intensive tests can be used instead (see Good, 1994). Frequently the description and comparison of interest in categorical data is simply a test of whether the proportion of some outcome of interest is the same in two samples; this can be done by a simple binomial test (see Fliess, 1981).

4.2.2. Ordinal Data

Comparison of ordinal data between two or more groups can be done by the same sort of $n \times m$ table methods described above for categorical data (and some ordinal extensions have been developed; see Agresti, 1984). Equally useful are rank-based techniques such as the Wilcoxon/Mann-Whitney and Kruskal-Wallis tests mentioned above.

A common comparative analysis performed on rating scale data is to look for improvements in ratings by comparing the means of two samples taken at different points in time, such as repeated surveys with different respondent samples. Even if calculating the mean for such a scale were reasonable (and it is for some ordinal scales whose behavior appears similar to ratio scales), the mean is sensitive to those few values at the skewed end which are of least interest. Thus any change in the mean at best only indirectly reflects the phenomenon of interest. Using the median does not have this problem, but suffers from the fact that the scale has few values and thus the median is likely to be the same from one sample to the next. There are two ways to compare such samples of rating scale data; both reduce the data to categorical data. The first method is to compare the entire distribution of responses across both samples in a $2 \times n$ table. The second method is to focus just on the category of greatest interest (say, the highest one or two), and compare the proportion of responses in that category in the two samples. While this method loses more information than the first, it focuses on the main area of interest and is easier to report and interpret.

4.3. Prediction

Frequently, measurements are made in order to predict the value of other measurements of interest. Such predictions do not have to be temporal ones; the notion of correlation is at bottom a predictive one: knowing the value of one measurement on

a unit, increases one's knowledge of the possible value of other measurements on it. The prototype of such prediction is regression. Originally limited to linear prediction equations and least-squares fitting methods, regression methodology has been extended over the course of the past century to cover an impressive variety of situations and methodologies using the framework of generalized linear models. Good references are Draper and Smith (1998), Rawlings et al. (1998), and Dobson (2001).

The essential method of regression is to fit an equation to pairs of measurements (X, Y) on a sample in such a way as to minimize the error in predicting one of the measures (Y) from the other (X). The simplest such case is where the regression equation is limited to a linear form:

$$Y = a + bX + \text{error}$$

and the total error measure is the sum of squared differences between the predicted and actual observations. The regression coefficient b then reflects the effect on Y of a 1-unit change in X. This notion of regression can then be generalized to prediction of a Y measure by a set of X measures; this is multiple or multivariate regression.

Even an elementary discussion of the method and application of regression is beyond the scope of this chapter (see Rosenberg, 2000 for one oriented toward software metrics), but a number of pitfalls should be mentioned.

First, most regression methods are parametric in nature and thus are sensitive to violations of their assumptions. Even in doing a simple univariate regression, one should always look at the data first. Figure 4 shows a cautionary example from Anscombe (1973); all four datasets have exactly the same regression line.

Second, regression models by definition fit an equation to all and only the data presented to them. In particular, while it is possible to substitute into the regression equation an X value outside the range of those used to originally fit the regression, there is no guarantee that the resulting predicted Y value will be appropriate. In effect, the procedure assumes that the relevant range of X values is present in the sample, and new X values will be within that range. This problem with *out of range* prediction complicates the use of regression methods for temporal predictions where the X value is time, and thus new observations are by definition out of range. For predicting temporal data, other methods must be used (as described in Sect. 5.3).

Third, regression equations have an estimation error attached to them just like any statistical estimate. Plotting the confidence bands around a regression line gives a good indication of how useful the equation really is.

Fourth, multivariate regression assumes that the multiple predictor measures are independent, i.e., uncorrelated with each other, otherwise the results will be incorrect. Since multiple measures are often correlated, it is critical to look at the pattern of correlations among the predictor variables before doing a multivariate regression. If even a moderate amount of correlation is present, something must be done about it, such as dropping or combining predictors.

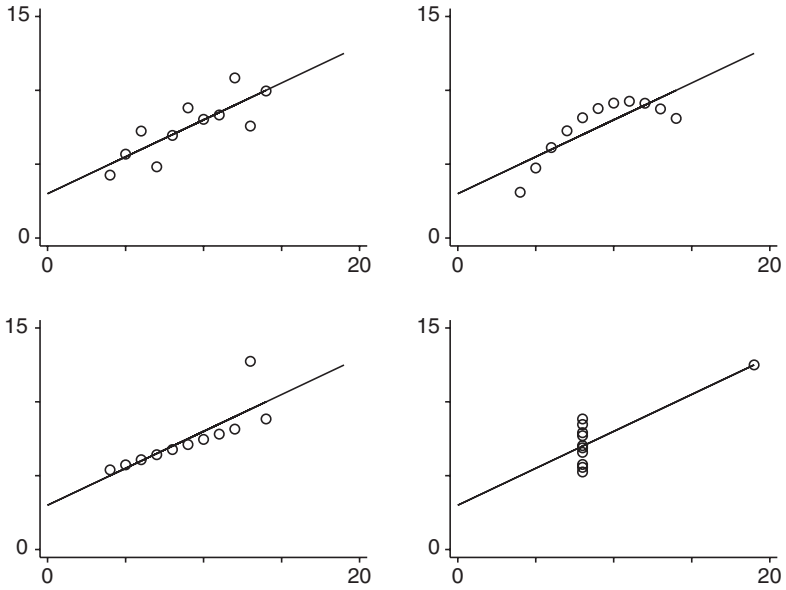

Fig. 4 Anscombe's example of four different data sets with exactly the same best-fitting regression line

4.3.1. Categorical Data

A frequent question of interest is how a binomial or other categorical variable can be predicted from another one, or from one or more ordinal or continuous variables (see El Emam et al., 1999 for an example in the area of software metrics). Such a prediction is sometimes called termed a *classification task*, especially if there are more than two categories; see Hand (1997) for a general discussion. The case of predicting a dichotomous outcome is termed a *diagnostic prediction* from its prototypical example in biostatistics: predicting whether or not a person has a disease based on one or more test outcomes. The accuracy in such a diagnostic situation can be characterized by a 2 × 2 table, as shown in Table 1, where the predictor variable(s) are constrained to make a binomial prediction which is then compared to the "true" value.[1]

Table 1. The structure of a prototypical diagnostic prediction

	Reality	
Prediction	Negative	Positive
Negative	True negative (A)	False negative (B)
Positive	False positive (C)	True positive (D)

[1] A known true value in such situations is called a *gold standard*; much work has been done on the problem of assessing predictive accuracy in the absence of such a standard (see, for example, Valenstein, 1990; Phelps and Huston, 1995).

Predictive accuracy in this context can be measured either as *positive predictive accuracy* (D/[C+D]), *negative predictive accuracy* (A/[A+B]), or both together (A+D/[A+B+C+D]). Two other relevant measures are *sensitivity*, the probability of correctly predicting a positive case, (D/[D+B]), and *specificity*, the probability of correctly predicting a negative case, (A/[A+C]).

There is an extensive literature on binomial prediction; much of it has been influenced by the theory of signal detection, which highlights a critical feature of such predictive situations: the prediction is based not only on the amount of information present, but also on some *decision criterion or cutoff point* on the predictor variable where the predicted outcome changes from one binomial value to the other. The choice of where to put the decision criterion inescapably involves a tradeoff between sensitivity and specificity. A consequence of this is that two prediction schemes can share the same data and informational component and yet have very different predictive accuracies if they use different decision criteria. Another way of putting this is that the values in any diagnostic 2 × 2 table are determined by both the data and a decision criterion. The merit of signal detection theory is that it provides an explicit framework for quantifying the effect of different decision criteria, as revealed in the *ROC* curve for a given predictive model, which plots the true-positive rate (sensitivity) and false-positive rate (1 – specificity) of the model for different values of the decision criterion (see Fig. 5). The ROC curve provides two useful pieces of information. First, the area under the curve above the diagonal line is a direct measure of the predictive accuracy of the model (the diagonal line indicates 50% accuracy or chance performance; a curve hugging the upper left

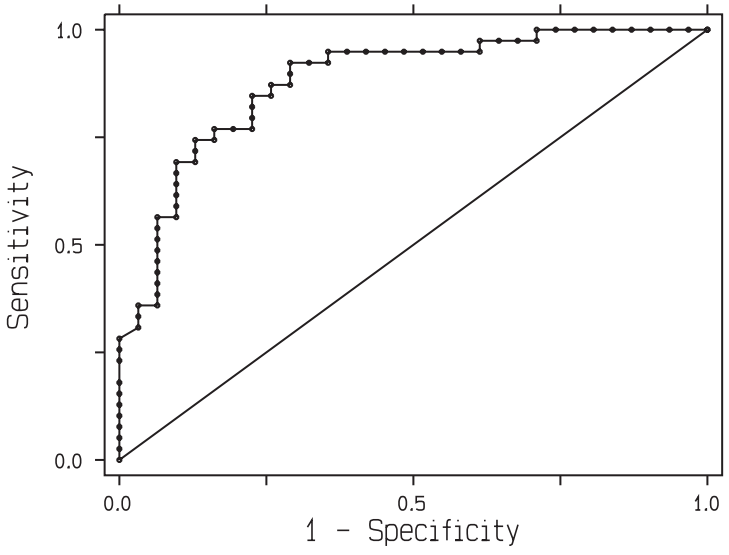

Fig. 5. An example receiver operating characteristic (ROC) curve

corner would indicate 100% accuracy). Second, one can graphically compare the relative accuracy of two models by their ROC curves: if the two curves do not intersect, then one model always dominates the other; if they do intersect, then one model will be more accurate for some values of the predictor variables. A good introduction to signal detection theory is Swets (1996). Zhou et al. (2002) provide a thorough guide to its application.

Regression methodology has been adapted for predicting binomial outcomes; the result is called *logistic regression* because the predictions have to be scaled by the logistic transformation so that they range between 0 and 1 (see Kleinbaum, 1994; Hosmer and Lemeshow, 1989). Coefficients in logistic regression have a somewhat different interpretation than in ordinary regression, due to the different context. The results of a logistic regression are often also expressed in terms of ROC curves.

4.3.2. Ordinal Data

Prediction of ordinal values is rarely done except by assuming that the values reflect an underlying interval or ratio scale, in which case standard regression methods are used.

5. Analyzing Dynamic Measurement Data

One of the most frequent uses of metrics is to track some attribute over time, either to detect or forecast changes in it, or to verify that the value is unchanging apart from unavoidable random variation. Such *time series data*, as they are called, have as their essential characteristic the presence of temporal structure. The chief structural patterns are *trend*, a long-term change in value, typically monotonic but sometimes cyclic in an aperiodic manner, or both; and *seasonal change*, a cycle of change with a fixed period, as with changes over the course of the seasons in a year. While the usual goal is to identify these temporal components, sometimes the goal is to demonstrate that no such components are present; such a time series is said to be *stationary*. It should be noted that analyses of time series data require at least three seasonal cycles worth of data, since estimating the seasonal component require more than one season's worth of data. Having less data seriously restricts the kinds of analyses that can be done, and usually arises in situations more accurately termed *longitudinal* or *repeated measures* analysis, where the goal is to examine relatively large-scale permanent changes such as physical growth or skill-acquisition. See Singer and Willet (2003) and Crowder and Hand (1990) for examples.

In addition to the methods described below, there are a great many other types of dynamic data analysis, such as survival analysis (mentioned briefly above), and state space models. See Gottman (1995) and Haccou and Meelis (1994) for examples.

5.1. Description

As with any analysis, the first step is to look at the data. Figure 6 shows a typical dataset containing a long-term increasing trend, with an additional seasonal component (every 12 months). The top panel shows the observed data, while the lower two panels display the underlying trend and seasonal components, respectively. Methods for such *time-series decomposition* are discussed in Bowerman and O'Connell (1993).

There are a number of ways such data can be used. The first way is simply to describe the history of some process. Rather than summarizing the history by a histogram or descriptive statistics such as the mean or standard deviation (which would miss entirely the temporal aspect of the data), the time chart and its decomposition into trend and seasonal components is the main focus.

Most discussions of time series analysis make the assumption that the observations are made with little or no error, otherwise the variation in the measurements themselves could obscure the temporal patterns. This means that this sort of analysis is best used on continuous measures (or counts) made with high reliability and precision, rather than ordinal measures such as ratings.

It is always important to verify that the temporal measurements in a time series are in fact equivalent. For example, fluctuations in the number of defects reported for each month in a 1-year period might seem to warrant some concern about quality variation, but in that respect they may be illusory. Months may seem equal, but they vary in length by up to 10%, and when the number of actual working days is

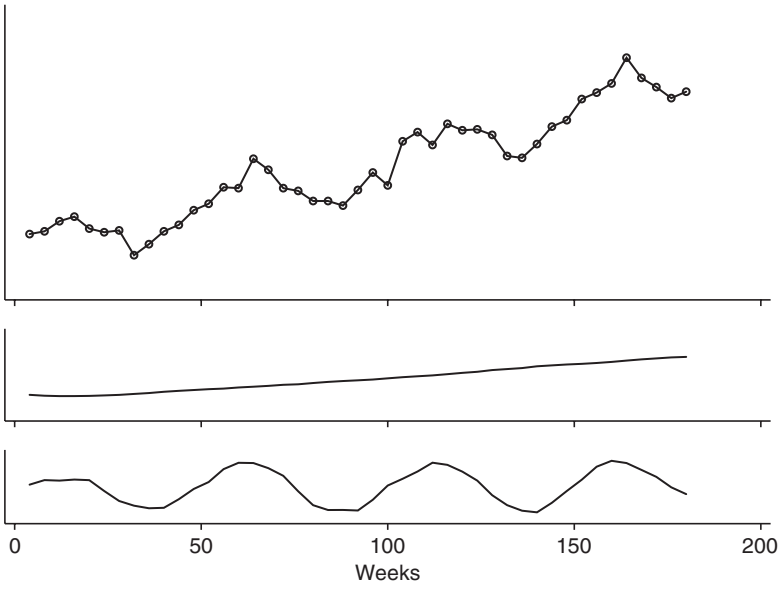

Fig. 6 Time series decomposition chart for data in Fig. 6

taken into account, they can vary by 25% or more. The same data adjusted for the number of work days may show little variation at all. This is not to say that the first approach is "false," merely that it can be seriously misleading if the variation in temporal units is not made clear. Even if the defect submission *rate* is constant from month the month, the actual *number of defects* submitted will vary; the first piece of information may be comforting for the quality manager, but the second piece is more valuable to the support manager.

5.2. Comparison

Often the question of interest is: "Is the latest observation evidence of a change in trend?" Such a question is difficult to answer on the basis of a single observation. Often, however, that observation is actually a summary of a number of observations, for example, the mean of some set of measurements. In that case one can use the same sort of statistical methods used with static data to compare the latest sample with the previous one. Typically, however, the sample sizes involved are too small to detect the small level of change involved. A more common method of looking for a change in trend is to compare the latest observation with the value predicted for it by a forecast.

5.3. Prediction

Another major use of time series data is *forecasting*: predicting one or more future observations based on the data at hand. The larger the amount of data at hand, the better the forecasting that can be done. Even with few data, however, there are some simple techniques that can be used. The simplest forecast technique is the so-called *naive predictor*, which assumes that the future value will be the same as the present value. This actually can be a useful first approximation in many cases, for example, tomorrow's temperature is likely to be similar to today's. Other naive predictors can be defined; for example, if there is a small amount of data beyond one seasonal cycle (say 15 months, January of one year to March of the following year) one can take the average difference between the observations made on the same part of the cycle (January to March for both years) and use that as an increment for forecasting the rest of second cycle based on corresponding values from the first.

Such naive predictors can be useful for first approximations, and can also serve as concrete points of departure for discussions about possible alternative forecasts. Perhaps most importantly, they can be used as baselines for evaluating the predictive accuracy of more sophisticated forecasting techniques.

There are a variety of ways of quantifying the accuracy of forecasts, all of them based on some measure of the difference between forecast and actual values. Chief among these are (here "error" and "deviation" mean the same thing):

Mean absolute deviation (MAD) the average absolute difference between observed and forecasted values (this penalizes errors in direct proportion to their size, and regardless of direction);

Mean squared error (MSE) the average squared difference between observed and forecasted values (this penalizes errors as the square of their size, also regardless of direction);

Mean percentage error (MPE) the average proportional difference between forecast and actual values (i.e., (*actual – forecast*/*actual*), expressed as a percentage;

Mean absolute percentage error (MAPE) the average absolute proportional difference, expressed as a percentage.

There are many more possible accuracy measures, each with its advantages and disadvantages; some may not be applicable with some kinds of data (for example, MPE and MAPE do not make sense when the data are not measured on a ratio scale with a zero point). Which to use depends on the purpose of the forecast, and which kinds of errors are considered worse than others (see Makridakis, 1998).[2]

Assessing the overall accuracy of a forecast is more complicated than in the case of static predictions with regression. A common technique is to set a desired standard of absolute or relative accuracy beforehand, and then compare the accuracy of various forecasting methods with that of a naive predictor. Often the choice of forecasting methods comes down to a trade-off between accuracy and difficulty of computation.

An additional issue to consider in forecasting is whether a forecast metric is a *leading*, *lagging*, or *coinciding indicator*, that is, whether changes in the metric occur before, after, or at the same time as changes in some other metric of interest. Leading indicators are highly desirable, but few metrics have that property. The issue is important because a metric cannot be effectively used for process control purposes unless its temporal connection with the process is understood.

5.4. Process Control

The other major use of dynamic, temporally oriented data is in determining that there is *not* change over time. This is the area of *statistical process control*.

A process is performing effectively if its behavior only changes under conscious direction; left alone it should remain stable, and measurements made on it should remain the same apart from the inevitable and unimportant random variation. In the 1920's Walter Shewhart at Western Electric devised a statistical method for quantifying and monitoring the stability of a process, the *control chart*, examples of which are shown in Fig. 7.

As can be seen, the control chart looks very much like a trend chart, except that it is based on a defined *control level* or expected value of the measurements (the

[2] These accuracy measures can also be used in assessing the fit of models to static data, of course, but in the latter case there are more useful global goodness-of-fit measures such as R^2 which are used instead. Such measures are not available for forecasting dynamic data.

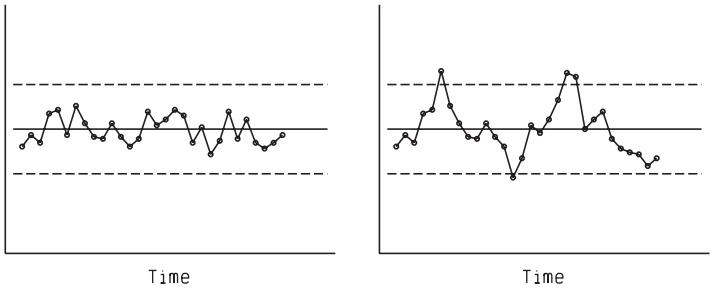

Fig. 7 Control charts showing (**a**) A process which is in control, (**b**) A process which is not in control

solid line), as well as *control limits* (the dashed lines), which define the range of values that are expected to be observed if the process is operating stably at the control level (and thus differences in observed measurements are due simply to random variation). There are different types of control chart, depending on the kind of measurement being tracked, such as continuous measures, counts, or proportions. Multivariate control charts track several measurements jointly. The overall principle is the same in each case: a baseline control level is established by a series of measurements of the process, and control limits are defined in terms of the observed variability of the process (and possibly also the desired variability). One then plots measurements of the process taken at regular intervals and looks either for measurements lying outside the control limits (and thus indicating that the process is operating outside of its normal range, presumably because of some interfering factor), or for patterns in the measurements which suggest that the observed variability is not random, but is due to some factor or factors affecting the process.

Figure 7a illustrates a process that is under statistical control; Fig. 7b shows one that is out of control and Fig. 8a shows one that, while apparently under control (being inside the control limits), shows patterns in the measurements that deserve investigation.

In the decades since they were first developed, there have been many different variations developed to handle the variety of process control situations that arise. One of the most useful variants is the *cumulative sum* or cusum chart, which is more sensitive at detecting changes in the level of process measurements. Cusum charts work by accumulating the deviations from the baseline expected value of the process; if the variation is truly random, the variations in one direction counterbalance those in the opposite direction and the cumulative sum remains close to zero. If, on the other hand, variations in the process are biased even slightly in one direction or the other, then the cumulative sum will advance towards the upper or lower control limit. This accumulation of small biases allows the trend to be detected earlier than would be the case with a standard control chart. Figure 8 shows both a standard chart and a cusum chart for a process that is drifting slowly out of control.

The theory and practice of control charts is highly developed and remains a central part of quality engineering. Good references are Montgomery (1996) and Duncan (1986). More recently, Box and Luceño (1997) have elaborated the relationship between statistical process control and engineering control theory.

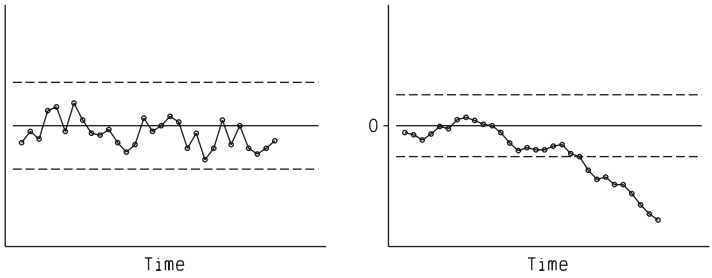

Fig. 8 A Process drifting slowly out of control as shown in (**a**) A standard control chart, (**b**) A cusum chart

There are also statistical methods for the optimization of process metrics, such as Evolutionary Operation (Box and Draper, 1969), response surface methodology (Montgomery and Myers, 2002), and data envelopment analysis/stochastic frontier analysis (Jacobs et al., 2006).

6. Data Quality

At this point, it is appropriate to return to the context of measurement and the dependence of statistical analysis on the quality of the underlying data collection process.

Data quality is a critical problem in industrial management, yet one often only vaguely recognized by decision makers who consume the ultimate endproducts of those data. This problem has come to light with the development of data warehouses, as warehouse developers discover that bad data can turn a data warehouse into a data garbage dump. The first step, then, in using measurements is ensuring that those measurements are of sufficient validity and accuracy to enable conclusions to be drawn from them.

The sources of data quality problems are manifold (apart from the question of bad metrics, dealt with in Sect. 3). Chief among them are

- Organizational problems
- Lack of precise definitions
- Lack of data validation
- Missing data
- Sampling bias

6.1. Organizational Problems

It is common for metrics to be defined and collected by people other than those to whom the metrics apply; this a recipe for trouble. The problem is exacerbated when a process is evaluated by management on the basis of metrics that the people carrying out the process find irrelevant or misguided; the inevitable result is distortion of

the work process to produce acceptable numbers, rather than valid or meaningful ones. For a metrics program to be successful, all parts of the organization involved need to be in agreement on the meaningfulness of the metrics and their role in the organization's effective functioning.

6.2. Lack of Precise Definitions

Many problems are caused by lack of a precise definition for a measurement. For example, measuring defects in software for whatever purpose, be it research or quality management, requires a clear definition of what constitutes a defect. This definition may reasonably vary depending on the question being asked (and the goal that question is answering), but whatever the purpose, the definition must address such issues as

- Are feature enhancement requests defects?
- Are usability problems defects?
- Are internally reported problems defects?

Similarly, measuring the time it takes to repair a defect requires addressing such issues as

- When does the clock start?
- Does it start at different times for internally vs. externally reported defects?
- When does the clock stop?
- What time is recorded if the repair of the defect turns out not to be a repair after all?

If these issues are not addressed at the time the metric is defined, then they will have to be addressed by those collecting the data if and when they arise. Not surprisingly, when that happens the results may not be as intended. The problem of vague definition is exacerbated when the measurements must be collected by different groups or individuals who often have, or develop over time, different interpretations of the definition. Such different definitions may go unnoticed for long periods of time until some situation brings it out.

Detecting the lack of precise definitions is done most directly by looking for explicit written documentation of what the definition of each of the measures is. In the frequent case where such information is lacking, it becomes necessary to interrogate those responsible for collecting, processing, and analyzing the data to find out what they have been assuming the measures' definitions to be; their answers will often be conflicting.

6.3. Lack of Data Validation

A precise definition for a metric is no guarantee that the values recorded for it make sense. It is very common to find observations with dubious or outright impossible values, due directly or indirectly to data-entry problems. These range from typing

errors to miscalibrated measuring devices to lack of understanding of the metric's definition. The presence of bad values is usually easy to detect if one takes the trouble to look; frequently, as long as the measurement process produces values that seem "reasonable" no-one bothers to audit the process to verify that the measurements are correct. For example, consider measurements of resolution times for customer problems that are derived from recording the dates and times when the service ticket is officially opened and closed. If there is no validation done to ensure that the closing time is chronologically later than the opening time, the derived resolution metric might take on zero or even negative values (perhaps from subtraction of a constant amount from all tickets; this would only become negative in ones with small values). Even if this occurs in only a small percentage of cases, it can seriously bias the estimates of resolution time. Simply dropping anomalous cases when they are found is not a solution until investigation has shown that such cases occur at random rather than for some systematic reason. Any particular case of bad data may have many potential causes which must be investigated; an occasional data entry error might be ignored, but a systematic distortion of entries cannot be.

Validation of data is the essential tedious first step of any data analysis. It can be made much easier and faster if the data are validated as they are collected. There are two difficulties which frequently prevent that from happening. First, those collecting the data are often not the ones who will use it for analysis, and thus have little understanding or interest in making sure that the data are correct. This is not due to maliciousness; it is simply due to different motivation. To take the above example, the people working the service desk have as their main goal the rapid processing of as many service tickets as possible; data validation interferes with this, with little or no visible benefit. Solving this problem requires educating management as well as the workers.

Second, even if validation is intended, it may be impossible to do in real time without degrading process performance. The general solution here is to arrange some way to do it "off line" rather than in real time, for example, validating new database entries overnight.

Detecting problems of data validation is done by performing extensive assertion- and consistency-checking of the dataset. For example, if the dataset contains measures of duration, they should be checked to make sure that each value is greater than zero. Often it is important to ensure that the value of one measure is logically compatible with that of some other measure. For example, a problem resolution of "replaced circuit board" is not consistent with a trouble report classified as "software problem."

6.4. Missing Data

It is rare to find a large dataset without missing values on at least some of its measurements, and care must be taken that missing-value codes (e.g., "99") are not mistakenly interpreted as genuine data values. (A particularly insidious case of this occurs with spreadsheets, which treat missing data as actually having the value "0.") This

raises the possibility that an analysis using only the available data may be subject to an unknown amount of error. The issues are therefore how much data can be missing without affecting the quality of the measurements, and what if anything can be done to remedy the situation. There is a large body of literature on this subject, which is discussed in the chapter by Audris Mockus in this volume.

6.5. Sampling Bias

The problems just discussed are easy to observe and understand. More subtle but just as serious is the problem of sampling bias. A precisely defined, thoroughly validated, complete dataset can still be useless if the measurement process only measures a particular subset of the population of interest. This can be for a number of reasons:

6.5.1. Self-selection

It may be that only some units in the population put themselves in the position of being measured. This is a typical problem in surveys, since typically there is little compulsion to respond, and so only those individuals who choose to be measured provide data. Similarly, only those customers with problems are observed by the customer service department.

6.5.2. Observability

Some measurements by definition are selective and can lead to subtle biases. For example, in a study of defect densities, some source modules will have no (known) defects and thus a defect density of zero. If these cases are excluded, then statements about correlates of defect density are true only of modules which have known defects, not all modules, and thus cannot easily be generalized. Another kind of observability problem can occur, not with the units being observed, but with the measuring device. For example, if problem resolutions are measured in days, then resolutions which are done in ten minutes are not accurately observed, since their time must be rounded down to zero or up to one day.

6.5.3. Non-random Sampling

A frequent problem in surveys, this also plagues many other kinds of measurements, including experiments where the selection of experimental units is not properly considered. Lack of information about the population, coupled with a bias to sample those units which are easy to sample, can result in a measured sample which is quite unrepresentative of the population of interest.

Detecting sampling bias can be difficult, because it typically happens before the data are collected. It can sometimes be spotted by the absence of certain kinds of data (customers from one region, service times longer than 1 month, etc.), but usually must be identified by studying the documentation for the data collection process or interrogating the people who carry it out. Correcting sampling bias is extremely difficult, since the basic problem is the complete lack of representation for some part of the population. To the extent that the type and degree of bias is known (also a difficult problem) it may be possible to adjust for it, but generally the only solution is to make it clear just what subset of the population is described in the dataset. A good discussion of detecting and coping with overt and hidden biases can be found in Rosenbaum (2002).

As should be clear from the above, problems of data quality are ubiquitous and difficult to deal with, particularly because there are only general guidelines for what to do, and each case must be handled on its own terms.

7. Summary

This chapter has discussed the role of the measurement process, the need for metrics to be clearly defined, reliable, and valid in order for them to be effective, and various statistical techniques and pitfalls in analyzing measurement data. Understanding measurement is a crucial part in the development of any branch of science (see Hand, 2004); the amount of effort devoted to it in empirical research in software engineering reflects the necessity of answering some of the most fundamental questions facing computer science and engineering. Fortunately, we can take advantage of the experience and knowledge gained by other disciplines, and apply them with advantage in developing effective software measurement.

References

Agresti, A, *Analysis of Ordinal Categorical Data*. New York: Wiley. 1984.

Agresti, A, *An Introduction to Categorical Data Analysis*. New York: Wiley. 1998.

Anscombe, F, Graphs in statistical analysis. *American Statistician*. 27(1):17–21. 1973.

Basili, V, Caldiera, G, and Rombach, D, The goal question metric approach. In: Marciniak, J, ed., *Encyclopedia of Software Engineering*. New York: Wiley. 1994.

Bevington, P, and Robinson, D, *Data Reduction and Error Analysis for the Physical Sciences*, 2nd ed. New York: McGraw-Hill. 1992.

Bowerman, B, and O'Connell, R, *Forecasting and Time Series: An Applied Approach*, 3rd. ed. Belmont, CA: Wadsworth. 1993.

Box, G and Draper, N, *Evolutionary Operation: A Statistical Method for Process Improvement*. New York: Wiley. 1969.

Box, G and Luceño, A, *Statistical Control by Monitoring and Feedback Adjustment*. New York: Wiley. 1997.

Briand, L, El Emam, K, and Morasca, S, On the application of measurement theory to software engineering. *Empirical Software Engineering*. 1(1). 1996.
Chayes, F, *Ratio Correlation*. Chicago: University of Chicago Press. 1971.
Cleveland, W, *The Elements of Graphing Data*. Summit, NJ: Hobart Press. 1994.
Cliff, N, What is and isn't measurement. In: Keren, G and Lewis, C, eds., *A Handbook For Data Analysis in the Behavioral Sciences*, Vol. 1: *Methodological Issues*. Hillsdale, NJ: Erlbaum. 1992.
Cohen, J, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillside, NJ: Erlbaum. 1988.
Comrey, A and Lee, H, *A First Course in Factor Analysis*, 2nd ed. Hillsdale, NJ: Erlbaum. 1992.
Crowder, M, and Hand, D, *Analysis of Repeated Measures*. New York: Chapman and Hall. 1990.
Dobson, A, *An Introduction to Generalized Linear Models*, 2nd ed. New York: Chapman and Hall/CRC. 2001.
Draper, N and Smith, H, *Applied Regression Analysis*, 2nd ed. New York: Wiley. 1998.
Duncan, A, *Quality Control and Industrial Statistics*, 5th ed. New York: Irwin. 1986.
El Emam, K, Benlarbi, S, and Goel, N, Comparing case-based reasoning classifiers for predicting high risk software components. National Research Council Canada technical report NRC 43602/ERB-1058. 1999.
Fenton, N and Pfleeger, S, *Software Metrics: A Rigorous and Practical Approach*, 2nd ed. Boston: PWS Publishing. 1997.
Fliess, J, *Statistical Methods for Rates and Proportions*, 2nd ed. New York: Wiley. 1981.
Ghiselli, E, Campbell, J, and Zedeck, S, *Measurement Theory for the Behavioral Sciences*. San Francisco: Freeman. 1981.
Good, P, *Permutation Tests*. New York: Springer. 1994.
Goodman, L and Kruskal, W, *Measures of Association for Cross Classifications*. New York: Springer. 1979.
Gottman, J, ed., *The Analysis of Change*. Hillsdale, NJ: Erlbaum. 1995.
Haccou, P, and Meelis, E, *Statistical Analysis of Behavioural Data: An Approach Based on Time-Structured Models*. Oxford: Oxford University Press. 1994.
Hand, D, *Construction and Assessment of Classification Rules*. New York: Wiley. 1997.
Hand, D, *Measurement Theory and Practice: The World through Quantification*. Oxford: Oxford University Press. 2004.
Hosmer, D and Lemeshow, S, *Applied Logistic Regression*. New York: Wiley. 1989.
Hosmer, D and Lemeshow, S, *Applied Survival Analysis*. New York: Wiley. 1999.
Jacobs, R, Smith, P, and Street, A, *Measuring Efficiency in Health Care: Analytic Techniques and Health Policy*. Cambridge: Cambridge University Press. 2006.
Keppel, G, *Design and Analysis: A Researcher's Handbook*, 3rd ed. New York: Prentice Hall. 1991.
Kleinbaum, D, *Logistic Regression*. New York: Springer. 1994.
Kleinbaum, D, *Survival Analysis*. New York: Springer. 1996.
Krantz, D, Luce, R, Suppes, P, and Tversky, A, *Foundations of Measurement*. New York: Academic. 1971.
Long, J, *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage. 1997.
Maddala, G, *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press. 1986.
Makridakis, S, Wheelwright, S, and Hyndman, R, *Forecasting: Methods and Applications*, 3rd ed. New York: Wiley. 1998.
Montgomery, D, *Introduction to Statistical Quality Control*, 3rd ed. New York: Wiley. 1996.
Montgomery, D and Myers, R, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 2nd ed. New York: Wiley. 2002.
Phelps, C, and Huston, A, Estimating diagnostic accuracy using a "fuzzy gold standard". *Medical Decision Making* 15:44–57. 1995.
Rawlings, J, Pantula, S, and Dickey, D, *Applied Regression Analysis*, 2nd ed. New York: Springer. 1998.
Rosenbaum, P, *Observational Studies*, 2nd ed. New York: Springer. 2002.

Rosenberg, J, A methodology for evaluating predictive metrics. In: Zelkowitz, M., ed., *Advances in Computers*, Vol. 23. New York: Academic. 2000.
Shepperd, M and Ince, D, *Derivation and Validation of Software Metrics*. Oxford: Clarendon Press. 1993.
Singer, J and Willett, J, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford: Oxford University Press. 2003.
Sprent, P, *Applied Non-Parametric Statistical Methods*, 2nd ed. New York: Chapman and Hall. 1993.
Swets, J, *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics*. Hillsdale, NJ: Erlbaum. 1996.
Taylor, J, *An Introduction to Error Analysis*, 2nd ed. Sausalito, CA: University Science Books. 1997.
Valenstein, P, Evaluating diagnostic tests with imperfect standards. *American Journal of Clinical Pathology* 93:252–258. 1990.
Velleman, P, Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician*. 47:65–72. 1993.
Wellek, S, *Testing Statistical Hypotheses of Equivalence*. New York: Chapman and Hall/CRC Press. 2002.
Wickens, T, *Multiway Contingency Tables Analysis for the Social Sciences*. Hillsdale, NJ: Erlbaum. 1989.
Zhou, X, Obuchowski, N, and McClish, D, *Statistical Methods in Diagnostic Medicine*. New York: Wiley. 2002.
Zuse, H, *Software Complexity: Measures and Methods*. New York: Walter de Gruyter. 1990.