6 Accurate observation

Doctors are trained to be careful observers. The teachings go back almost 2500 years to Hippocrates of Cos, who wrote:

It is the business of the physician to know, in the first place, things similar and things dissimilar; those connected with things most important, most easily known, and in anywise known; which are to be seen, touched, and heard; which are to be perceived in the sight, and the touch, and the hearing, and the nose, and the tongue, and the understanding; which are to be known by all the means we know other things.

Hippocrates on the appearance of the face in impending death

(Hippocratic Facies)
'... [the physician] should observe thus in acute diseases; first, the countenance of the patient, if it be like those of persons in health, and more so, if like itself, for this is the best of all; whereas the most opposite to it is the worst, such as the following; a sharp nose, hollow eyes, collapsed temples; the ears cold, contracted, and their lobes turned out; the skin about the forehead being rough, distended and parched; the color of the whole face being green, black, livid, or lead-colored. If the countenance be such at the commencement of the disease, and if this cannot be accounted for from the other symptoms, inquiry must be made whether the patient has long wanted sleep; whether his bowels have been very loose; and whether he has suffered from want of food; and if any of these causes be confessed to, the danger is to be reckoned so far less; it becomes obvious, in the course of a day and a night, whether or not the appearance of the countenance proceeded from these causes. But if none of these be said to exist, and if the symptoms do not subside in the aforesaid time, it is to be known for certain that death is at hand.'

The Hippocratic dicta led to the development of medical semeiology, the study of the signs and symptoms of disease. Medical practitioners began to compare, in detail, the condition of disease with that of health. In traumatic injuries, it became the practice of ancient surgeons to compare the injured part very carefully with its corresponding part on the opposite side. And 'all the means [by which] we know other things' referred to a new method

of reasoning that replaced abstract speculation and the proclamation of oracles. The physician was encouraged to consider the logical consequences that follow from the information provided by the senses.

An original observation concerning the urine of a patient afflicted with diabetes as reported to the Medical Society of London in 1776 by Matthew Dobson

'Experiment V

After evaporating two quarts of urine to dryness by gentle heat, there remained a white cake, which was granulated and broke easily between the fingers. It smelled like brown sugar, neither could it from the taste be distinguished from sugar.'

Beginning with Hippocrates, attention was focused on the patient; the shift to accurate observation in medicine was as fundamental as the one that occurred when astrology was transformed into astronomy.

Doctor Joseph Bell's observations and deductions

In *The Man who was Sherlock Holmes*, Hardwick and Hardwick have written of the time when the young medical student, Arthur Conan Doyle, was appointed as out-patient clerk to his teacher, Doctor Bell. Sitting back in his chair, the surgeon quickly noted the peculiarities of the patients who were ushered into his room by Doyle. And he would address his clerk and a circle of medical students as follows: 'Gentlemen, I am not quite sure whether this man is a cork cutter or a slater. I observe a slight callus, or hardening, on one side of his forefinger, and a little thickening on the outside of this thumb, and that is a sure sign he is either one or the other.' Another case was simple: 'I see you're suffering from drink. You even carry a flask in the inside breast pocket of your coat.' A third patient listened open mouthed as Bell, after saying 'A cobbler, I see,' turned to his students and pointed out that the inside of the knee of the man's trousers was worn; that was where the man had rested the lapstone, a peculiarity only found in cobblers. One example of Bell's diagnoses impressed Doyle so much that he never forgot it:

'Well, my man, you've served in the army.'
'Aye, sir.'
'Not long discharged?'
'No, sir.'
'A Highland regiment?'
'Aye, sir.'
'A non-com. officer?'
'Aye, sir.'
'Stationed at Barbados?'
'Aye, sir.'

'You see, gentlemen,' explained Bell to his students, 'the man was a respectful man but he did not remove his hat. They do not in the army, but he would have learned civilian ways had he been long discharged. He had an air of authority and he is obviously Scottish. As to Barbados, his complaint is of elephantiasis, which is West Indian and not British.'

PHYSICIAN AS DETECTIVE

In the modern era, the name Sherlock Holmes has become the embodiment of the power of reasoning from observed facts. It is of interest that Holmes' creator, Arthur Conan Doyle, began his career as a physician and that the fictional detective was modeled after one of his teachers in medical school. The prototype, Doctor Joseph Bell, a Scottish surgeon at the Edinburgh Infirmary, explained to Doyle and to other medical students:

The precise and intelligent recognition and appreciation of minor differences is the real essential factor in all successful diagnosis [...] Eyes and ears which can see and hear, memory to record at once and to recall at pleasure the impression of the senses, and an imagination capable of weaving a theory or piecing together a broken chain or unraveling a tangled clue, such are the implements of his trade to a successful diagnostician.

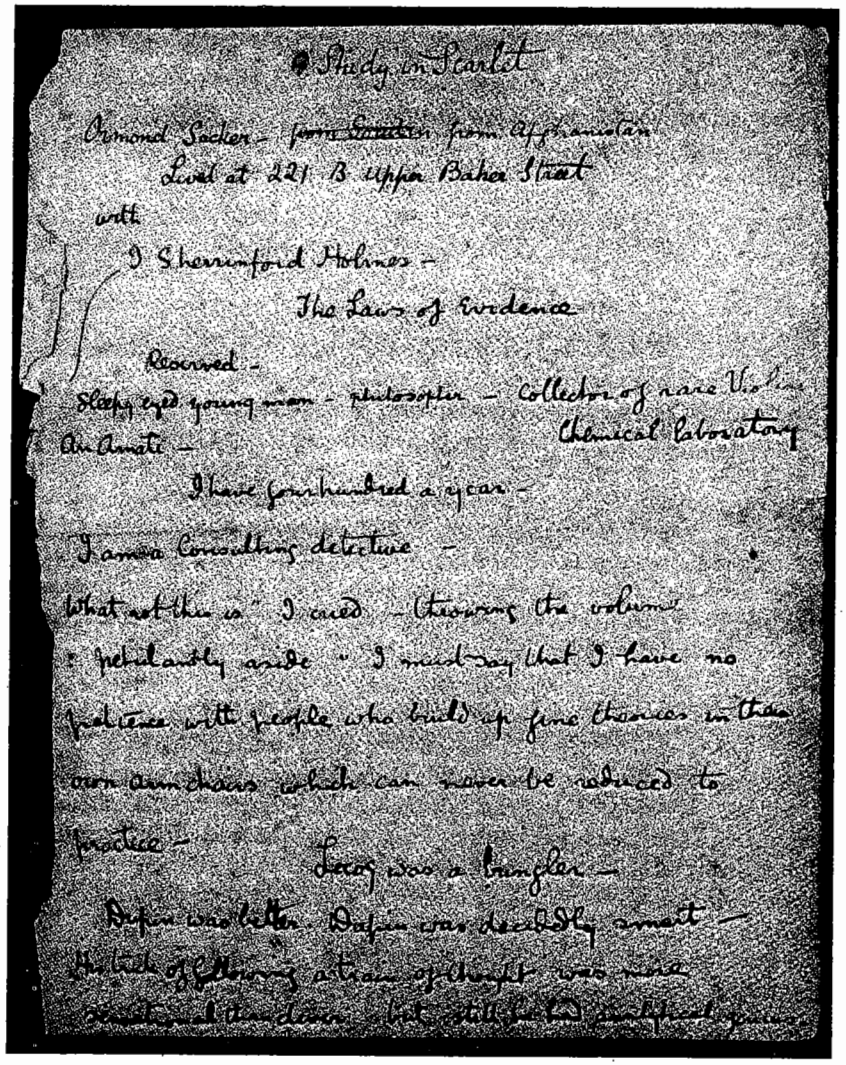
Although Doyle's stories celebrate the power of deductive reasoning, the detective's methods were not strictly deductive. Martin Gardner of *Scientific American* has noted that Holmes first tried to gather all the evidence that was relevant to the problem at hand, like a scientist trying to solve a mystery of nature. At times, he performed experiments to obtain fresh data. He then surveyed the total evidence in the light of his vast knowledge of crime, and of sciences relevant to crime, to arrive at the most probable hypothesis. Then the theory was further tested against new evidence, revised if need be, until the highly probable 'truth' emerged triumphantly.

The prepared mind

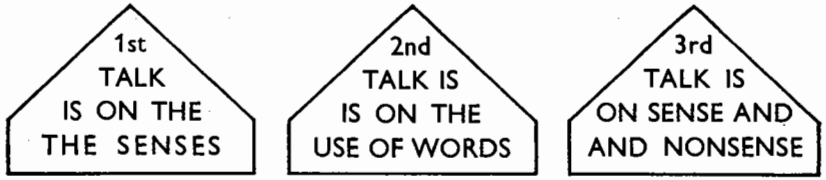
Notice, once more (p 15), that the sleuth's *initial* observations were, to use Popper's phrase, theory-impregnated. Holmes examined the evidence with a question in mind. The master detective's apotheosis, 'Joe' Bell, noted details about patients that entirely escaped the medical students (whose visual acuity was, very likely, much better than that of their older role model) because the experienced observer 'looked' with a prepared rather than an open mind. Moreover, the teacher used a limited number of 'off-the-shelf' theories; they were not constructed *de novo*. The inside of a trouser leg, frayed in a certain way, indicated to Bell the effect of a lap-held stone on which shoemakers beat leather, not the limitless possibilities that might be entertained by his students even after the physical clue was pointed out to them.

Pattern recognition methodology (used by detectives and by physicians) is, of course, a useful aid in everyday work but limitations do arise in the presence of novelty: we tend to see things we anticipate rather than the things that are there.

Preliminary notes on the character of Sherlock Holmes



In the first conception of the detective (who was to appear in the story *A Study in Scarlet*), Arthur Conan Doyle noted these words to himself: 'The Laws of Evidence.'



The lantern slide announcing the topics to be discussed in Asher's Lettsomian Lectures contains three printing errors.

Prefigured observation

The English clinician, Richard Asher, demonstrated our propensity to make what might be called prefigured observations and our unconscious dismissal of the anomalous. At the beginning of his Lettsomian Lectures delivered before the Medical Society of London in 1959, Asher projected a lantern slide announcing the topics to be covered in the series of talks. He intentionally 'planted' three gross printing errors in the announcement and later pointed out to the audience that only one person had caught the slips. A physician is both blessed and cursed by this suppressive mechanism, he noted, blessed when he detects an expected and significant pattern but cursed when significant irrelevancies are set aside without being appreciated.

The hidden man

This incomplete picture demonstrates how a significant pattern remains hidden until the relevant and irrelevant features are sorted out. Once these are pointed out (see the 'solution' on p 76), it becomes almost impossible to *disregard* the face in this picture.

The recognition of significant patterns depends on past experiences and education, Asher argued, and he projected a figure of The Hidden Man (p 75) to demonstrate the point. Before the solution is 'taught', the pattern seems meaningless; after seeing the completed figure (below), the eye finds it almost impossible to return to its previous state, which was inexperienced, untutored, and, as a consequence, relatively unbiased.

'Solution' to the hidden man

Additional details are added to the picture on p 75 to make the man's head and shoulders obvious. Now the pattern in the incomplete drawing cannot be erased from the 'seeing eye.'

PROPHECY ON TRIAL

It is unintentional 'within-observer' predilection, not conscious forgery, that is implied by the term 'observer's bias' in clinical trials involving conscientious physicians seeking fair appraisals of proposed treatments. The problem of biased observation must not be ignored; as I have emphasized repeatedly, observers should be 'masked' whenever it is technically possible and practical to do so.

On the surface, it may seem that scepticism is carried to unreasonable lengths when it is transferred from a concern with biasing effects of largely impersonal forces in the case of 'passive' observation to suspicions about the observer himself in 'active' observations. It may also appear frivolous to call for the judgment of an 'innocent' observer in the manner of a magician who must try to convince an audience of his supernatural powers.

> We travel like Ulysses to learn about the particolored world and its motley ways, but what we see is mostly controlled by patterns formed in our minds long before we took the first step.
>
> Robert M. Adams

These tactics seem out of place in a medical setting and when applied to matters of life and death. Indeed, both physicians and patients often resent the machinations; the precautions are perceived as an unworthy intrusion on a relationship of mutual trust. But when all is said and done in critical tests of hypotheses in medicine, acts of prophecy are on trial and prophets are simply not credible witnesses under these circumstances.

MEASUREMENT

When a proposition is subjected to a formal trial, the terms used in the measurement of the question to be addressed need to be specified. Before continuing, it will be useful to consider what is meant here by the word *measure*, since the concept of measurement is broadened in medicine to include more than the procedures used in physics to determined length, time, and mass. In the less precise operations ordinarily used in the social sciences and in medicine, several levels of measurement need to be distinguished.

Nominative classification

Nominal scales of measurement are used in the basic, descriptive stage of development of any scientific study. Persons or objects are classified with respect to a specified characteristic that can be identified reliably; but the characteristic does not have the property of size. For instance, when numbers are assigned to individuals according to country of birth (1. American, 2. Canadian, 3. French, ...) these numbers of a nominative scale cannot be ranked by size in a graded sequence. The aim is to sort individuals or other elements into defined categories; all individuals assigned to the category are equivalent in terms of the characteristic.

An axis of classification is particularly useful in medicine when collections of individuals are assembled who are alike with respect to some illness-related qualities. In the national RLF study, for example, the enrolled infants were sorted into 'single birth' or 'plural birth' classes; this identified groups of babies with strikingly different frequencies of RLF (twin, and other multiple births were affected three times more frequently than singletons). Classification is the *sine qua non* of all organized study of natural events; it is self-evident that all other levels of measurement, no matter how precise, involve categorization as a minimal operation. Moreover, if the

classes are exhaustive (for instance, 'single birth' and 'plural birth' categories include all new born infants) and mutually exclusive (each baby can belong to only one class), we have the minimum requirements for numerical analysis. Notice that the numbers used to count the frequency of individuals assigned to the subclasses reflect the operation of enumeration, not the operation of nominal scaling.

Exclusionary effect of categorization Although nominative classification is a powerful first step in organizing a body of information or an orderly search for new information, it does have shortcomings. Whenever data are summarized, certain information is inevitably lost. As a result, classification limits and defines the range of possible conclusions that can be extracted from the descriptive facts. The quip 'What you say what you get' aptly describes this constrictive effect of classification. Labeling bias stipulates the level of entry into an inquiry, and such effects cannot be completely eliminated by refined definitions. The nominalists are quite correct when they maintain that to define *is* to leave out. It must be recognized that each summarizing process is an abstraction which imposes limitations, and the implications that flow from the level of abstraction used must not be ignored.

Problem-specific classification In recent years, classification of infants by birthweight *and* by functional indicators of development has provided considerable insight into problems that were blurred previously, but the categories are still quite non-specific. RLF occurs almost exclusively in infants with incomplete development of the blood vessels of the retina; thus, an elementary 'one-zero' scale for the 'presence' or 'absence' of an immature network of retinal vessels would be a useful first step in a problem-specific classification of babies for RLF studies. (This variety of unranked scale is called 'existential'.) Unfortunately, technical barriers make it impractical to use the criterion in the smallest babies as the normally clear, forward portion of the eye—the vitreous—is hazy in the first days of life and this makes it difficult to inspect the retinal layer which lies behind it, until the infants are one to two weeks old.

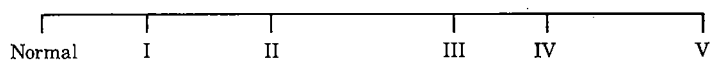
Practical difficulties of this kind are common in medicine: the realities must be acknowledged, but they are also quite challenging. Formal investigations often provide an opportunity for the development of improved nominal scales for the highly relevant but vaguely defined information collected by observation at the bedside. The physician observer is the 'measuring instrument' and the object is to devise classifications that are directly related to clinical problems.

Ordinal succession

The level of measurement is raised when we can make use of the concept of numbered succession in the form of ordinal scales. This refinement becomes possible when the magnitude of qualities can be compared. If we can place individuals in an ordered succession, along a continuous scale of size, there will be 'ties' and 'near ties' who resemble one another very closely and they may be assigned to mutually exclusive classes. But the between-class distinctions are now quantified. This kind of measurement is higher than the primary level; for, in addition to the nominative operation, the categories can be ranked in a size-determined order.

In RLF, the blood vessels of the retina undergo changes that can be detected with an ophthalmoscope. The aberrations begin when infants are several weeks old and progress over a period of weeks and months; the changes may regress completely, go on to leave scars in the eye, or advance to complete blindness. Stages of increasing severity were classified (in 1953) into ten grades (stages I to V for the early blood vessel changes and a set of five grades for the late scarring lesions). These ordinal scales were very useful in making semiquantitative statements about RLF when the disorder was occurring in epidemic frequency.

The ordinal level of measurement does not supply any information about the size of differences between the grades of the continuum. We can only declare (in RLF) that the larger numbers represent an increased intensity of the process, but we cannot express the relationships in familiar mathematical terms. Addition or subtraction is possible, but only in a restricted sense. For instance, if RLF activity (early blood vessel changes) is represented by distance along a continuum:



we can say that the distance I-V = I-II + II-III + III-IV + IV-V, but we cannot make valid statements about the relationships between distances I-II and II-III ... IV-V.

When we translate order relationships with indeterminate intervals into arithmetic operations, we cannot justify the use of the ordinary processes of addition, subtraction, multiplication, and division. Nonetheless, the fact that we can make comparisons such as 'greater than' and 'less than' represents a step up from the lower scale of measurement.

Scores Dimensional appraisal of the physical signs and subjective symptoms of illness may be expressed in the form of scores. Severity or intensity, for instance, are ranked according to specified rules for assigning one, and only one, number on an ordinal scale of values. The score is developed by choosing a number of items that are thought to be related to the quality

that is characterized and then selecting, by trial, the elements that are consistent in the way they order patients.

Although relatively few of these scalar methods have been used in medicine, there is no logical reason for avoiding these aids to the organization of the rich store of observations about the uniquely human aspects of illness.

Interval and ratio scales

The next levels in the hierarchy of measurements (interval and ratio scales) require the use of standard units (such as meters, kilograms, and degrees of temperature) to express quantity. A distinction is made between an interval system in which a numbered scale of equal units begins at an arbitrary zero and a ratio system of numbers in which zero indicates absence of a measured property. An interval scale allows us to determine only that two objects differ by a certain amount of the property. The highest scale of measurement permits statements about the ratios that obtain between the measured properties.

For example, the Celsius system of temperature measurement is an interval scale (0°C is arbitrarily set at the freezing point of water); as a result we can determine, say, that $50°C - 25°C = 25°C$, but $50°C \div 25°C \neq 2$; we are not justified in declaring that 50°C is twice as warm as 25°C. Temperature measured in the Kelvin system is a ratio scale (O°K, absolute zero, is the temperature at which molecules stop moving and there is literally 'no heat'); here differences *and* ratios reflect the real world. There are no restrictions on arithmetic operations carried out on numbers obtained in ratio scale measurements.

Valid arithmetic

Notice that each level of measurement has distinct properties, and that the levels themselves form a cumulative scale. An ordinal scale has all the characteristics of a nominal scale plus ordinality. An interval scale has all the qualities of both lower scales plus a unit of measurement, and a ratio scale, representing the highest level, has a unit of measurement and a meaningful zero. It is logical to drop back one or more levels if necessary in the course of analyzing data, but it is quite illogical to move up the hierarchy if the underlying assumptions necessary to perform arithmetic operations have not been satisfied.

There are no inherent safeguards that prevent us, for example, from using numbers obtained in ordinal measurements for a calculation which requires a higher scale (such as calculating the arithmetic mean of the stages of RLF); only reasoning back to the nature of the measurement will prevent such errors. The appropriate use of measurement, like the correct use of language, depends entirely on the user.

Addition

A man begging for money summarized his plight on a sign next to his tin cup. It read:

Wars	2
Legs	1
Wives	2
Children	4
Bankruptcies	2
Total	11

ACCURACY IN MEASUREMENT

There are no units of measure for many of the signs and symptoms that constitute the raw knowledge of bedside medicine. As a result, laboratory measurements that can be expressed by interval and ratio scale numbers are often collected as proxies for ordinal or nominal information in the hope of reducing subjective influences on observations.

Observer error

There is a mistaken belief that rigor and precision are associated only with the higher levels of measurement. An uncritical sense of confidence seems to develop when measurements are made by others using instruments at some distance from the bedside. I find it interesting that the observers who are entirely dependent on instruments to make measurements on the objects farthest removed from our direct sense experience, the astronomers, were the first to become curious about the personal element in instrumental errors of measurement.

Kinnebrook's defect In August 1795, Maskelyne, the royal astronomer at the Greenwich Observatory, found that his assistant, Kinnebrook, was recording the movements of stars across the sky about a half-second 'too slow' (when compared with Maskelyne's records). He was convinced that all through 1794 there had been no discrepancy. Maskelyne cautioned Kinnebrook about the 'error'; nevertheless, it increased during the succeeding months until, in January 1796, it had become about eight tenths of a second. At this point, Maskelyne dismissed his assistant.

The error was serious; the calibration of the Greenwich clock, the standard of time on which many other calculations were made, depended upon these stellar transit speeds. (They were made by an 'eye and ear' method: the field of a telescope was divided by parallel crosswires in the reticle; the observational problem consisted of noting, to one tenth of a second from the audible beats of a clock, the time at which a given star crossed a given wire.) Kinnebrook's error of eight tenths of a second was a relatively large

one and seemed to justify Maskelyne's conclusion that the man had fallen 'into some irregular and confused method of his own'.

The incident was recorded in the pages of *Astronomical Observations at Greenwich*, and would have passed into oblivion had it not come to the attention of an astronomer at Königsberg, named Bessel, some twenty years later. The experience struck him as odd and he considered the possibility that the errors were not willful. It seemed to Bessel that Kinnebrook, when informed of his 'error', must have tried to correct it. The failure to succeed, he thought, might mean the error was involuntary. Bessel set out to study the observations of stellar transits made by a number of senior astronomers. Differences in observation, he discovered, were the rule, not the exception.

'Personal equation' of observers A good deal of interest in the problem was generated by these findings and efforts were made to 'calibrate the observer' in the hope of correcting for the deviations of observation. In the 1840s, the practice of measuring the 'personal equation', as it came to be called, became common among astronomers. In addition to determining the 'absolute' personal equation for each observer, the 'personalities' of the eye, of the ear, and of touch were calculated. The prevailing notion was that these variations between observers were related only to physiological differences (and were unique for each astronomer), but, by the 1870s, it slowly became clear that there were psychological variants that accounted for the personal variability in observations.

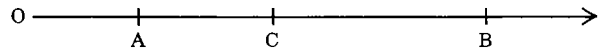
Confirmation bias Contemporary research on the measurement performance of scientists has demonstrated that observer errors tend to produce results which lean in the direction of the observer's hypothesis. In one set of studies, recording errors by experimenters in behavioral research were not frequent (1 per cent of over 20 000 observations), but when the errors did occur, more than two thirds of the time they were in the direction of the recorder's hypothesis. Laboratory workers, looking for hours at a pointer on a scale or digital displays, develop subconscious ideas, it seems, as to the 'proper behavior' of the inanimate devices.

Properties of sets of measurements

The third replicate Readings in a time sequence often reveal a type of expectancy bias and an effort must be made to ensure that repeat measurements are truly independent. The issue arises, for example, when three successive measurements of a single object are made to calculate the mean. If the last value does not fall in the range between the first two, there is a temptation to replace it with a 'better' measurement. Many years ago, W.J. Youden of the National Bureau of Standards pointed out that chance

The behavior of sets of measurements

Consider a set of three measurements which are made sequentially, Youden has said. After the first two are made, how often will the third measurement fall between the two? (No assumption needs to be made that the measurements follow a symmetrical distribution.) Indicate on a scale the position of three measurements, A, B, and C:



Here C does lie between A and B, but the three measurements might have fallen in any one of six ranked orders in which they can be arranged:

Rank order	*1*	*2*	*3*
	A	C	B*
	A	B	C
	B	A	C
	B	C	A*
	C	A	B
	C	B	A

*These are the only two, out of six equally likely sequences, that bring C between A and B

Thus, it follows that one third of the sets of measurements will be such that the third measurement falls between those already made. Two times out of three, in the long run, the third measurement will be smaller than both or larger than both the first pair. The general formula for this property of data reads: If $(n-1)$ measurments are followed by an nth measurement, the chance that this measurement falls between the smallest and the largest of the $(n-1)$ measurements is $(n-2)/n$. For example, in sets of ten measurements, once out of five times, the tenth measurement will be either smaller or larger than the other nine.

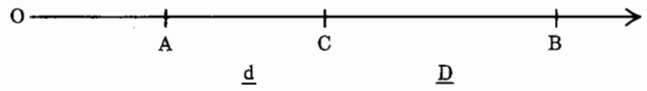
variation of independent measurements is worth considering when these are made in sets.

The behavior of sets of three measurements is such that the third measurement falls between the two already made only one third of the time, in the long run. This property of data cannot be changed by anything the observer can do, since it results from the action of the goddess of fortune.

Size of the outlying replicate School courses in quantitative chemical analysis often require that students turn in duplicate analytic results. The grade for the exercise depends in part on how well the average of the two determinations agrees with the value ascribed to the material and, in part, on how well the two agree with each other. Students commonly perform three 'runs' and turn in the pair showing the best agreement. Studies of the behavior of measurements, however, indicate that the dispersion of the average of selected pairs is greater than that exhibited by unselected duplicates. The properties of sets of three measurements are such that it appears

The 'outlying' value in sets of three measurements

When a set of three measurements is made, it frequently happens that two are in close agreement and the third lies considerably removed from the pair. There is a strong temptation to discard the outlying value, on the basis of an intuitive feeling that a blunder accounts for the apparent discrepancy. Unless there is some conception of the ways three measurements may distribute themselves when there is nothing whatsoever wrong with any of them, a judgment as to whether or not a 'slip' has been made is almost certain to err in discarding good values. Consider again, Youden argued, the three measurements marked off on a linear scale; this time the distances between them are denoted as d and D:



It has been shown that the interval D is at least four times as large as d more than a third of the time, and it is ten or more times as large as d in 15.7 per cent of sets of three measurements. Values of the ratio D/d exceed 32.57 once in twenty times according to studies conducted at the National Bureau of Standards. Thus, if there is no knowledge about the dispersion of values, there is little justification for discarding the remote value unless it is removed from the closest pair by an amount some thirty times the difference between the measurements constituting the closest pair.

unwise to discard an 'outlying' value unless it is removed from the closest pair by an amount some thirty times the difference between the close-lying pair.

Terminology

Terms to distinguish the kinds of errors encountered in observations are used rather loosely. To reduce confusion, it is helpful to define the words when they are used in reports of medical studies.

The term *accuracy* is often used to indicate the degree of agreement between an observation and its true value; the latter can be determined when refined methods of measurement are available. Similarly, measurements are said to be *unbiased* if they do not systematically overstate or understate the true value of a characteristic. On the other hand, *precision* (*reliability* and *consistency*) measures the extent to which a series of observations agree with one another—the repeatability of the results of measurement.

The distinction between accuracy (lack of bias) and precision (reliability, consistency) is of importance in clinical medicine, where there is often preoccupation with repeatability because there is no objective method of measuring the 'true' value. A set of bedside observations can be inaccurate

yet quite precise. Agreement between observers, and in repeated observations by the same observer, gives no assurance of accuracy.

> Immersion in water makes the straight seem bent: but reason thus confused by false appearances is beautifully restored by measuring, numbering and weighing: these drive vague notions of greater or less or more or heavier right out of the minds of the surveyor, the computer, and the clerk of the scales. Surely it is the better part of thought that relies on measurement and calculation.
>
> Socrates

Validity of measurement tends to connote accuracy, but, as with the latter term, the exact meaning is fuzzy when there is no accepted outside standard for comparison. For example, validity is often defined as the extent to which an observer measures what he or she claims to measure. The definition is suspended in thin air if the claim is described in terms of the measure: 'intelligence is what intelligence tests measure.' I will explore these problems further in the next chapter.

Observations of natural events proceed, it has been said, with a 'mixture of clear logic and unwritten superstition'. And yet the replacement of hunches and guesses in medicine with measurements of the course of disease and the effects of treatment is secure. Measurement is the fibre of modern medicine.