STATISTICS - AN OVERVIEW

BY

Erich L. Lehmann

Research Supported by
National Science Foundation Grant No. DMS84-03188.

Technical Report No. 72
July 1986

Department of Statistics
University of California
Berkeley, California

E.L. Lehmann

STATISTICS - AN OVERVIEW

Contents

Introduction.

The term "statistics" is used in different ways. "Accident statistics", or "Sales statistics" refer to numerical data in these areas. (The corresponding theoretical terminology defines statistics to be any functions of observable random variables). Common problems encountered in the work with such data and those collected by scientists, engineers, government officials, lawyers, doctors,... have led to the development of general methods and principles concerning the collection, presentation, and analysis of data. The term statistics is also used as the discipline concerned with such methods.

The present article considers statistics broadly as the field comprising all of the above concerns. It might be described as the enterprise dealing with the collection of data sets, and extracting and presenting the information they contain.

Comprehensive surveys of the present state of the field of statistics are provided by the nine volumes of this Encyclopedia and the two volumes of the International Encyclopedia of Statistics. (Reference to articles in the first of these will be indicated by an * and in the latter by a +. To locate these articles, it may sometimes be necessary to consult the index of the Encyclopedia.) However, like any other scientific disciplines, statistics is an ongoing, everchanging endeavor. This is obvious for the collection of data relating to constantly changing populations. Regularly published guides to the enormous volume of government statistics are provided by the American Statistical Index, The Statistical Abstracts of the United States, and the Statistical Reference Index (for the U.S.), and by the Index to International Statistics. Many other types of data are covered by

Statistical Sources: A Subject Guide to Data on Industrial, Business, Social, Educational, Financial, and Other Topics for the U.S. and Internationally, which is also periodically updated.

On the methodological side, new methods and formulations are constantly being developed, new types of application come into view and in turn give rise to new problems. This flow of research is disseminated through a multitude of journals being published around the world. A listing of annual contributions (more than 8000 in 1984) is available in the Current Index to Statistics: Applications, Methods, and Theory, published since 1975.

What establishes statistics as a discipline is that the same kind of data requiring the same kind of concepts and methods turn up in many different fields. On the other hand, specific areas of application also may require some specialization and adaptation to particular needs. Separate articles in this Encyclopedia outline how statistics is used in Actuarial work, Agriculture, Animal science, Anthropology, Archaeology, Auditing, Crystallography, Demography, Dentistry, Ecology, Econometrics, Education, Election forecasting and projection, Engineering, Epidemiology, Finance, Fisheries research, Gambling, Genetics, Geography, Geology, Historical Studies, Human genetics, Hydrology, Industry, Law, Library science, Linguistics, Literature, Manpower planning, Management science, Marketing, Medical diagnosis, Meteorology, Military science, Nuclear material safeguards, Ophthalmology, Parapsychology, Pharmaceutics, Physics, Psychology, Public administration and policy, Quality control, Sociology, Survey sampling Taxonomy, and Zoology.

I. Data Interpretation

Statistical Methodology.

It is a central fact, underlying essentially all statistical thinking, that an actual data set typically is only one of many possible such sets that might have been obtained under the given circumstances. (For a possible exception, see Diaconis [12].) Measurements vary when they are repeated. A store inventory, besides depending on the day on which it is taken, will be affected by bookkeeping and counting errors. In a survey, different households would be obtained if a new sample were drawn; even if the same households are visited on another occasion, different members may be at home and provide different answers; and, finally, even the same family member may answer the same questions differently on another day.

As a consequence, interpretation of a data set depends not only on the actual data but also on what (if anything) is assumed about the possible alternative observations that might have been obtained instead. The following sections consider three categories of such assumptions, and briefly indicate the kinds of statistical procedures that can be based on each.

Data Analysis.

It is rare that a data set is studied without any preconceived notions. Consider however the idealized approach of pure data analysis in which the data are considered on their own terms. The statistical methods developed on this basis have as their primary aim

(i) exploration of the data to uncover the features of principal interest,

and

(ii) presentation of the data in a manner that will bring out and highlight these features.

The set of techniques dealing with (i) and (ii) are called exploratory data analysis* and descriptive statistics+ respectively. Both employ a great variety of numerical and graphical*+ techniques.

The simplest, most basic data analytic methods concern a single batch of numbers, for example, the first-year sales of the twelve novels of a successful author, forty measurements of corrosion taken at different locations on a copper plate, or the ages of 350,000 cancer patients listed in a tumor registry. Histograms*, stem and leaf displays, and one-dimensional scatter plots* are some of the many ways of presenting the numbers of a batch. From these one can get an impression of where the data are centered (the general level of their values), and how spread out they are. Observations that lie far from the bulk of the data, socalled outliers ,*- may correspond to errors or exceptional cases and may deserve special attention. The display may exhibit unusual features such as bimodality, perhaps suggesting the possibility of a mixture of two batches, each unimodal but with different modes*. If there is marked asymmetry, the possibility arises of making a transformation* (such as taking logarithms) to obtain a more symmetric data set.

Instead of a fairly detailed display of the data, authors frequently present only one or two summary statistics; for example, the mean* or median to indicate the general level of the numbers, and a measure of their variability such as the standard deviation*, median absolute deviation, or interquartile range*. A compromise is the five number summary*, consisting of the smallest, largest, and median observation, and the first and third quartile. These may be displayed graphically as a box plot. (See Notched Box and Whisker Plots*).

Additional information concerning the numbers of a batch may be available and important. If the order is known in which the twelve novels appeared, the data may show that the sales steadily increased, or that they increased up to a certain point and then leveled off, and so on, thus providing an indication of the author's changing reputation and success.

Example 1. Corrosion Data. The exploratory use to which even very simple data can be put is illustrated by 40 corrosion readings taken at random over a metal plate (Campbell [6], Wolfowitz [36]). No particular pattern emerged when the observations were plotted according to their position on the plate. However, when they were plotted in the order in which they were taken, a bunching together in runs* of high and low observations suggested (as it turned out, correctly) a malfunctioning of the delicate measuring device.

Much of data analysis deals with batches of more complex units such as pairs of numbers, more general vectors, matrices, or curves, and it need not be restricted to a single batch. With multivariate data, one may be interested in the separate features of the different variables, in relationships among the variables or sets of variables, or in aspects of the overall pattern of the multidimensional data set. The development of graphical (including in particular computer-displayed) and numerical methods for these purposes is a very active area of statistical research. (See Multivariate graphics*.) It includes, among others, such approaches as cluster* analysis, and multivariate classification+, multidimensional scaling*, pattern recognition*, and factor analysis*+. When several batches are being considered simultaneously, comparisons of the batches will tend to be of primary interest.

Statistical analysis based on probability models.

The data-analytic approach indicated in the preceding section can provide clarification of the phenomena represented by both simple and very complex data sets, and can lead to important new insights and hypotheses. In its simplest forms such as numerical summaries and histograms, it is the statistical presentation most frequently encountered by the general public in newspaper articles and magazine reports. (Through misleading use, it also lends itself to much mischief. See for example. Huff [24].) However, this approach lacks what is often an essential requirement of the resulting inferences: Because of the fact, mentioned earlier, that the same phenomena might have led to different observations, it is impossible to assess the reliability of the conclusions.

Such as assessment requires knowledge concerning the alternative data sets that might have been observed in the given situation instead of the set that was observed. The crucial step underlying the modern theories of statistical inference and decision making is to consider the observed data as realizations of random variables. The possible values of these variables are governed by probability distributions specifying the probabilities of observing the various possible data sets. The unknown aspect of the situation (for the clarification of which the data were collected), is represented by the fact that we don't consider the probability distribution to be known, but only assume it to belong to a postulated family of possible distributions.

For a single batch $X_1, \cdots, X_n$, for example a set of n measurements of some quantity, investigators often assume that the n observations are independent, and that each has the same probability distribution. If this common distribution is denoted by F, the model is completed by specifying a family $\mathbb{F}$ to which F is assumed to belong. This may for example be the family of all possible distributions F (i.e. no further assumptions are made), or the family of all distributions which are symmetric with respect to a specified or unspecified point of symmetry. Such broad families are called nonparametric+ in distinction to parametric families where F is known except for the values of some parameters, for example, the family of all normal, Poisson, or Weibull distributions. Once the model is specified, one can now ask, and answer, the type of question concerning a single batch considered in the preceding section, with more precision. In particular, the conclusions no longer refer to this particular data set, but to the underlying process that produced it.

Suppose for example that the aspect of interest is the overall level previously represented by the average of the observations. Suppose that $\mathbb{F}$ is the family of normal distributions with mean ξ and unit variance. Then ξ is the average value not of these particular n measurements but rather of all potential measurements, each weighted according to its probability. The average $\bar{X} = (X_1 + \cdots + X_n)/n$, which before was a descriptive measure of the general level of the batch, now becomes an estimator of the unknown ξ: for example, the true value of the quantity being measured or the average value of the characteristic (e.g. height, age, or income) in the population from which the X's were obtained as a sample.

The model assumptions make it possible to get an idea of the accuracy of an estimator such as $\bar{X}$, for example in terms of the expected closeness to the true value ξ. One commonly used measure of this closeness is the expected squared error, which for the case of n independent measurements with variance 1 equals

$$E(\bar{X} - \xi)^2 = 1/n.$$

This formula shows for instance how the accuracy improves with n, and enables one to determine the sample size n required to achieve a given accuracy.

This approach also provides a basis for comparing the accuracy of competing estimators. For the median $\tilde{X}$ of the X's for example, one finds that approximately (if n is not too small) $E(\tilde{X} - \xi)^2 = 1.57/n$, more than 50% larger than the corresponding value $1/n$ for $\bar{X}$. The median is thus considerably less accurate than the mean. This conclusion depends strongly on the assumption that the X's are normally distributed. For other distributions the result may be just the reverse. (See for example the section on robustness in the article on Estimation*).

Another way of describing the accuracy of $\bar{X}$ is obtained by noting that

$$(1) \qquad P[(|\bar{X} - \xi|) \leq 1.96/\sqrt{n}] = .95,$$

so that with probability .95 the estimator $\bar{X}$ will differ from the true value ξ by less than $1.96/\sqrt{n}$. The statement (1) can be paraphrased by saying that the random interval $(\bar{X} - 1.96/\sqrt{n}, \bar{X} + 1.96/\sqrt{n})$ will contain the unknown ξ with probability .95. Random intervals that cover an unknown parameter ξ with probability greater than or equal to some prescribed value γ are called confidence intervals*+ at confidence level γ.

Point estimation and estimation by confidence intervals provide two of the classical approaches to statistical inference. The third is hypothesis testing*+.

Example 2. Extrasensory Perception. Suppose the claim of a subject A to have extrasensory perception (ESP) is to be tested by tossing a coin 100 times at a location invisible to A, and recording A's perception (Heads or Tails) for each toss. Suppose A obtains the correct result on 54 of the tosses. This clearly is not an indication of a strong ability at ESP. However, even a very slight ability would be of extraordinary interest. Is there support for such a finding, or is the result compatible with purely random guessing? Under the null hypothesis* of pure guessing, the probability of calling a toss correctly is 1/2 for each toss, and the probability of getting 54 or more right is then $\doteq 1/4$. The result is therefore not particularly surprising under H, and a case for A's ability to do better than pure chance has not been made. Had A correctly identified 65 tosses rather than 54, the conclusion would be quite different. The probability of 65 or more correct calls is only .002 when H is true. In the light of so extreme a result, one would have to give serious attention to A's claim.

Hypothesis testing, and point and interval estimation, are all used extensively in applications, with each being more useful and popular in some areas than in others. The theory of these three methodologies is concerned with the performance of proposed procedures (including the determination of sample size to achieve a desired performance), the comparison of different procedures, and the determination of optimal ones. An important consideration is the robustness* of a given procedure under violation of the model assumptions. If the procedures are very sensitive to these assumptions, one may want to study the problem in a nonparametric setting of the kind described for a single batch at the beginning of this subsection. Nonparametric (and semiparametric*) models are particularly important for the analysis of large data sets, which has become more practicable as a result of increased computer capabilities and availability.

A unified framework for the three areas is provided by Wald's Decision Theory*. This very general theoretical approach deals with the choice of one of a set of possible decisions d on the basis of observations x_i let the chosen d be denoted by $\delta(\underset{\sim}{x})$. The observed value $\underset{\sim}{x}$ is assumed to be the realization of a random quantity $\underset{\sim}{X}$ with probability density $p_\theta(\underset{\sim}{x})$, θ unknown. A loss function $L(\theta,d)$ measures the loss resulting from decision d when θ is the true parameter value, and the performance of a decision procedure δ is measured by its risk function

$$R(\theta,\delta) = E_\theta[L(\theta,\delta(\underset{\sim}{X}))],$$

the average loss incurred by δ when θ is true. A principal concern of the theory is the determination of a δ for which the risk function is as small as possible in some suitable sense. Another problem is the characterization of all admissible* procedures, i.e. all procedures whose risk cannot be uniformly improved.

Decision theory can also be made to encompass Sequential Analysis*, the Design of Experiments (by letting the loss function take account of the cost of observations), and the choice of model (by imposing a penalty that increases with the complexity of the model). However, it is an abstract approach which has been useful primarily for exhibiting general relationships rather than for its impact on specific methods. An important consequence of Wald's theory has been its liberating effect on the formal consideration of new types of situations, among them multiple comparisons and other simultaneous inference procedures.

Exploration vs Verification.

To illustrate both the relation and the difference between the approaches described in the preceding two sections, consider once more the ESP example.

Example 2. Extrasensory Perception (continued). When looking over the results of the 100 tosses, suppose the experimenter notices that of the 54 successes, 33 occurred during the last and only 21 during the first 50 tosses. The probability of 33 or more successes in 50 tosses with success probability $p = 1/2$ is only about .01. One might be tempted to explain away the poor performance in the first half of the experiment by the theory that it requires some warming up before the ability hits its stride, and to declare the success of the second half significant. However, such a conclusion would not be justified on the basis of this analysis since the calculation does not take account of the fact that the particular test (restricting oneself to the second half) was not originally planned but was suggested by the data.

Suppose the situation had been reversed, that there had been 33 successes in the first and 21 in the second half. A possible explanation: the exercise of ESP requires great concentration, and after 50 attempts the subject is likely to get tired. Similar explanations could have been found if success had been unusually high in the middle 50, or on the last 25, or the first 25, and so on. Instead of high concentration of successes in a particular segment of the sequence, other patterns might have struck an observer: for example, a gradual rise in the frequency of success, or a cyclic pattern of successes and failures, and so on. For each, an explanation could have been found.

This is the problem of multiplicity. Every set of observations - even a completely random one - will show some special features, and explanations can usually be found post facto to account for them. Unfettered examination of many different aspects of a data set is legitimate, and in fact a primary purpose, of exploratory data analysis. However, the results will then tend to look more significant than they really are since attention is likely to focus on the extremes of a large number of possibilities. To legitimize a theory suggested by the data one must test it, for example, on a separate part of the data not used at the exploratory stage or from new data specially obtained for this purpose. (A somewhat more limiting alternative to such a two-stage procedure is to formulate a number of possible theories to be considered before any observations are taken. The simultaneous testing* of such a number of possibilities provides a legitimate calculation for the probability of the most striking of the associated results.)

The two stages: exploration of the data leading to the formulation of a hypothesis, followed by an independent test of this hypothesis, constitute the basic pattern of scientific progress as described by scientists (see for example Feynman [14], Chapter 7) and discussed by philosophers of science. Before considering a third aspect in the next section, let us briefly mention another distinction, which relates to the purpose of a statistical investigation. This is the difference between inference* and decision making. It may be illustrated by the problem faced by a doctor who wants to arrive at a diagnosis (inference) but must also select a treatment (decision).

Example 3. Medical Diagnosis. Suppose there are k possible conditions (diagnoses) $\theta_1, \cdots, \theta_k$ that might have led to the observed symptoms and test results, $\underset{\sim}{X}$ (the data), including for example measurements of temperature, blood-pressure, and so on. Under condition θ, the observations $\underset{\sim}{X}$ have a distribution $p_\theta(\underset{\sim}{x})$. The problem of diagnosis is thus the statistical problem of using the observed value of x to determine the correct value of θ. (A standard procedure is to select the value $\hat{\theta}$ of θ that maximizes $p_\theta(\underset{\sim}{x})$ for the given observation $\underset{\sim}{x}$, the socalled maximum likelihood estimate*+ of θ). The associated decision procedure might be to select the treatment that would be most appropriate for the chosen θ. The situation is however more complicated since the choice of treatment must also take account of the severity of the consequences in case of an incorrect diagnosis resulting in an nonoptimal treatment, the socalled loss function. The distinction between inference and decision making is reviewed in Barnett [2]. For a discussion of computer-based medical diagnosis and treatment choice, see Shortliffe [31], and Spiegelhalter and Knill-Jones [34].

Bayesian inference and decision making.

Example 3. Medical Diagnosis (continued). Suppose a patient P. being tested for the conditions $\theta_1, \cdots, \theta_k$ of the last example is told that the tests point to $\hat{\theta}(= \theta_1$, say) as the most liely cause of the symptoms. Not unnaturally, P. wants to know just how likely it is that θ_1 is in fact the true cause. The doctor has to admit that the term 'most likely' was used imprecisely; that $\hat{\theta} = \theta_1$ is the condition which assigns the highest probability to the observed test results, not necessarily the most likely of the conditions $\theta_1, \cdots, \theta_k$, and that in fact no probability can be assigned to these conditions.

Actually, in this example it may be possible to make such an assignment. Suppose that π_i is the incidence of condition θ_i in the population of sufferers from the given symptoms, and hence is the probability that a patient drawn at random from the population of such sufferers has condition θ_i The condition of such a patient is then a random variable Θ, with prior probability $P(\Theta = \theta_i) = \pi_i$ before the tests are taken, and posterior probability $P(\Theta = \theta_i \mid \underset{\sim}{x})$ in the light of the test results $\underset{\sim}{x}$. The latter probability can be calculated from the π_i and the $p_{\theta_i}(\underset{\sim}{x})$ by Bayes' theorem*.

On being presented with this probability, P. may however still not be satisfied but complain to the doctor: "You have treated me for 20 years, you know my complete medical history, life style, and habits. In the light of all this additional information, what is the probability of suffering from θ_i not for a random patient but for me personally?". Unfortunately, the interpretation of probability as frequency, which was tacitly assumed up to now, and which in particular applied to the prior probabilities π_i of the preceding paragraph, precludes assigning probabilities (other than 0 or 1) to unique events such as this particular patient's suffering from condition θ_i.

This difficulty is met head-on by the Bayesian approach according to which π_i can be chosen to represent the physician's probability that condition θ_i obtains for this particular patient. The meaning of probability is however different from the earlier one. Probability is no longer a frequency but the degree of belief* attached to the event in question. (If the event is repetitive, this probability typically draws close to the observed frequency as the number of cases gets large.)

In general, the Bayesian approach assigns a <u>prior probability distribution</u>* to a parameter θ before the observations $\underset{\sim}{X}$ are taken. Once the values $\underset{\sim}{x}$ of $\underset{\sim}{X}$ are available, the prior distribution of θ is updated to the posterior (i.e. conditional) distribution of θ given $\underset{\sim}{X} = \underset{\sim}{x}$, which shows how the prior beliefs regarding the chances of different θ-values have changed in the light of the data.

In the decision theoretic terminology introduced earlier, the relevant assessment of the performance of a procedure δ from a Bayesian point of view is not the risk function $R(\Theta,\delta)$ but rather the posterior expected loss

$$r(\underset{\sim}{x},\delta) = E[L(\theta,\delta(\underset{\sim}{x}) \mid \underset{\sim}{x}],$$

calculated according to the conditional distribution of θ given $\underset{\sim}{x}$.

Ideally, a Bayesian has a comprehensive, consistent view of the world with a probability attached to every unknown fact. These probabilities must satisfy an appealing set of axioms (the axioms of <u>coherence</u>*), and must be updated as new information becomes available. For an individual's response to the world (or even a specific problem), a chief difficulty in implementing this program is the determination of the prior distribution*. A considerable literature deals with methods for eliciting a person's degrees of belief* with respect to a given situation.[1]

Ideally, each person has a single correct personal probability regarding any unknown event. However, in practice, these probabilities "can never be quantified or elicited exactly (i.e. without error) especially in a finite amount of time" (Berger, [3] p. 64). This has led to the suggestion of a robust Bayesian viewpoint according to which one "should strive for Bayesian behavior which is satisfactory for all prior distributions which remain plausible after the prior elicitation process has been terminated" (Berger, l.c.).

A second difficulty arises in situations in which the decision or opinion concerns not a single person, but represents a joint problem for a group or occurs in

[1] There is also an objective Bayesian school in which the prior distribution represents total lack of information. This will not be discussed here.

the public domain, as is the case for example in the publication of the analysis of a scientific investigation. Some aspects of this problem will be considered in the next section.

The Bayesian/Frequentist controversy.

The mutual criticism of Bayesians and Frequentists[1] has given rise to a lively (and sometimes acrimonious) debate, which has helped to clarify a number of basic statistical issues. One of the central concerns is that of subjective versus objective data evaluation in scientific inference and reporting. Fisher*, Neyman*, and E.S. Pearson* developed their frequentist theories in a deliberate effort to free statistics from the Bayesian dependence on a prior distribution, and this aspect has continued as the central frequentist objection. The Bayesian response to this criticism is twofold.

On the one hand, it is pointed out that frequentist analysis involves similar types of specification. There is the choice of model and loss function, both of which must be chosen in the light of previous experience and involve judgments which are likely to vary from one person to another. In addition, there is the problem of selecting a frame of reference which forms the basis of the frequency calculations. In assessing the incidence of the conditions $\theta_1, \cdots, \theta_k$ of the preceding section, for example, for what population should this be calculated: the population of the world, or the country or city in which the person lives; should the comparison be restricted to patients of the same age, sex,,...? This is

[1] There are many different variants of Bayesians and Frequentists, not all of which will agree with the positions ascribed to these approaches here.

the problem of conditional inference*, which so far has found no satisfactory frequentist solution. (For the Bayesian, the problem does not arise in this form since the probabilities will always refer to this particular patient, but the same issue arises when one must decide how to weigh the experience with other patients in forming an opinion about this one).

As a more positive response, there have been recent Bayesian suggestions (for example, Dickey [13], Smith [33]), that in scientific inference the analysis should be reported under a variety of different priors which - it is hoped - will include the opinions of the readers. The view that "any approach to scientific inference which seeks to legitimize an answer to complex uncertainty is a totalitarian parody of a would-be rational human learning process" (Smith [33]), considerably narrows the gulf between the two approaches.

The reason for this narrowing can be found in the combination of two facts. The first is a basic result of Wald's (frequentist) decision theory to the effect that every admissible procedure is a Bayes solution or a limit of Bayes solutions. Secondly, frequentists tend not to believe in a unique correct approach, and may therefore try a number of different solutions corresponding to different optimality principles, robustness properties, and so on. If these lead to similar conclusions, any one of them can be adopted. Otherwise, a careful examination of the differences may clarify the reason for the discrepancies and point to one as the most appropriate. Lacking such a resolution one may instead prefer to report a number of different procedures.

The Bayesian and frequency approaches lead to different ways of assessing the performance of a decision procedure. From a strict Bayesian point of view, only the posterior distribution of θ given $\underset{\sim}{x}$, and the posterior expected loss $r(\underset{\sim}{x},\delta)$, are relevant, while frequentists measure the performance of a procedure by its risk function. However, on this issue also, an accommodation to statistical practice and the need for communication has narrowed the gap by generating a Bayesian interest in risk functions. Thus Rubin [30], p.1161) writes: "Frequency calculations that investigate the operating characteristic* of Bayesian procedures are relevant and justifiable for a Bayesian when investigating or recommending procedures for general consumption". Similar considerations can be found in Berger [4].

In the other direction, the frequentist approach has been strongly influenced by Bayesian ideas, in particular, by recognizing that it is natural and useful to consider the prior distribution leading to a proposed admissible procedure. An example in which such a Bayesian interpretation has made an important contribution to a theory developed in a decision theoretic framework is that of Stein estimation*.

II. Data Acquisition

Measuring single units.

Part I dealt with the interpretation of data once they have been collected. In this and the following sections, we consider some of the processes that produce data, and the statistical problems arising at this earlier stage. (An extensive discussion of various types of data is provided in Hoaglin et al [21].)

The basic data units are the numbers, symbols, words, or other entries making up a data set; for example, measurements of the height of a person or plant, or of the weight of a wagonload of fruit or a minute amount of some chemical; barometer or temperature readings; or the scores of a student on an intelligence or aptitude test. Alternatively, the observations could be the information provided by a person answering a questionnaire or interviewer such as family size, last year's income, or religious and political affiliation.

A concern for data quality, for their reliability and validity, is an important task preceding the collection of data. Efforts must be made to eliminate bias*, reduce variability, and eliminate sources of error. In constructing a questionnaire, great care is required to avoid ambiguities. Checks on the reliability of responses can often be built into the data set, for example by asking for the same information in a number of different ways or in different contexts.

Another aspect of data improvement arises after the data have been obtained. Data editing* (or "cleaning") involves the deletion or modification of entries which do not appear to be in consonance with the rest of the data and which sometimes represents obvious errors (for example, in a series of monthly measurements of a baby's head circumference when one month's measurement is smaller than the preceding ones). A variety of statistical methods have been developed for this purpose. In addition, robust statistical procedures are available, which satisfactorily control the effect of outlying observations. (See for example Hoaglin et al [22] and Hampel [18]). On the other hand, data cleaning by inspection, without clearly stated rules, runs the risk of introducing a subjective element. (For example, just which observations are to be singled out for such treatment?). Even with good rules, it may remove valuable evidence and destroy the basis for probability calculations. For this reason, it is typically better to consider the editing of data as part of the statistical analysis and, while perhaps indicating definite or suspected errors, not to change the original data.

The improvement of data quality mentioned above is not the only aspect of data collection to be considered before obtaining the observations for an investigation. Two questions in particular, which are of crucial importance and which will be discussed in the next sections are how many, and what kind of observations are required.

Assessing population characteristics.

The target of most investigations is not a single unit but a population of such units. A set of professors, apples, days, mice, or light bulbs, is typically examined not because of interest in these particular specimens but in order to reach conclusions about the populations they represent. (It is this aspect and the associated quantification that has earned statistics its reputation of being antihumanistic).

Efforts to obtain information about every member of the target population through a census*+ go back to at least the third millenium B.C. in Babylonia, China, and Egypt. (For a nontechnical history of the census, see Alterman[1].) The purpose was usually to provide a basis for taxation or proscription. Today, population censuses seeking a great variety of information are carried out, many of them at regular intervals, in nearly all countries of the world.

However, taking a complete inventory is not the only way, and often not the best way, to obtain accurate statistical information about a population. The same end can usually be achieved more economically by collecting information only from the members of a sample, taken from the population by a suitable sampling* method. The much smaller size of the sampling operation tends to make this procedure not only cheaper but also more accurate since it permits better control of the whole process and hence the quality of the resulting data.

On the other hand, a census has the advantage - not shared by any sample - of providing information even for very small subpopulations.

Suppose a population Π consist of N units, each of which has a value v of some characteristic of interest, such as the age or income in case of a human population, the number or weight of the apples on each of the N trees in an orchard, or the length of life or brightness of the lightbulbs in a shipment received from the factory. To obtain an estimate of the average v-value $\bar{v} = (v_1 + \cdots + v_N)/N$, a sample is taken from the population. In the early days of sampling, investigators usually relied on judgment samples, in which judgment is used to obtain a sample 'representative'* of the population. It is now realized that such sampling tends to lead to biases, and does not provide a basis for calculating accuracy or the sample size needed to achieve a desired accuracy. The methods used instead are probability sampling schemes* which select the units to be included in the sample according to stated probabilities. The simplest such scheme is simple random sampling according to which n units are chosen in such a way that every possible sample of size n has the same probability of being drawn. The average of the sample v-values is then the natural estimate of $\bar{v}$.

Better accuracy can often be attained, and the needed sample size and resulting cost therefore reduced, by dividing the population to be sampled into strata within which the v-values are more homogeneous than in the population as a whole but which differ widely among each other. (The population of school children of a city might for example be stratified by school, grade, and gender.) A stratified* sample of size n is then obtained by drawing a simple random sample of size n_i from the i-th stratum for each i, where $\Sigma n_i = n$.

In a different direction, the cost of sampling can often be reduced by combining units into clusters* (for example, all the apartments in an apartment house, all houses in a city block, or all patients in a hospital ward), and obtaining the required information for each member of a sampled cluster. The two approaches can be combined into stratified cluster sampling, and many other designs are possible. For references to the extensive and sophisticated methodology available, see the articles on sampling*+ and sample surveys*+.

Sample surveys to obtain information about a population are in widespread use and have become familiar through election polls, market surveys, and surveys of television viewers to establish ratings. However, they are not very well understood. In particular, it appears puzzling how it is possible to obtain an accurate estimate of the opinions or intentions of many millions from information concerning just a few thousand.

Example 4. An election poll. To get some insight into this question, let us simplify the situation, and suppose that a population Π consists of a large number N of voters, each of whom supports either candidate A or B. A simple random sample of n voters is drawn from Π, and the preference of each member of the sample is ascertained. If X is the number of voters in the sample favoring A, then X/n, the proportion of A-supporters in the sample, is the natural estimate of the proportion p of A-supporters in the population.

The question at issue is how large a sample is required for X/n to achieve a prescribed accuracy as an estimate of p, for example, for the standard deviation (SD) of X/n to satisfy

$$\text{(1)} \qquad SD(X/n) \leq .01.$$

It is often felt intuitively that the required sample size n should be roughly proportional to the population size N. However, it turns out that this intuition is misleading and that in fact for large populations the required n is essentially independent of the value of N.

To see this, suppose for a moment that the sampling is done "with replacement", i.e., that the members are drawn successively at random, with each - after giving the required information - being put back into the population before the next member is drawn, again at random. This method is slightly less efficient than the original simple random sampling (and therefore requires a larger sample size), because it allows the same unit to be drawn more than once. It is introduced here because it is particularly simple to analyze. Sampling with replacement is characterized by two properties: (i) On each draw, the probability of obtaining an A-supporter is p; (ii) the results of the n draws are independent in the sense that the probability of getting an A-supporter on any given draw does not depend on the results of the earlier draws.

Let us now compare this situation with a quite different one. Suppose a coin with probability p of falling heads when spun on its edge (this probability may be far from 1/2), is spun n times and that X is the number of times it falls heads. Then (i) the probability of heads is p on each spin, and (ii) the results of the n spins are independent. The standard deviation of X/n is therefore the same for this coin problem as in the election-sampling with replacement. The number n required to reduce this standard deviation to .01 is therefore also the same in both cases. Note however that the coin problem involves only n and p; no population is involved. Therefore the required n cannot depend on N.

This argument depends of course crucially on the assumption that the sampling was random. It is the randomness which insures that with high probability the sample contains approximately the same proportion of A-supporters as the population. In addition, it was tacitly assumed that the preference of each voter can be ascertained without error. In practice, the possibility of "response error" can rarely be ruled out.

Data from experiments.

A study is called an *experiment* if its data are produced by an intervention (i.e. do not occur naturally) for the purpose of gathering information. It is a *comparative experiment* if its purpose is to compare several ways of doing something (e.g. different teaching methods, medical treatments, fertilizers, and so on) rather than to determine some absolute value.

Example 5. Weather modification*. Consider a company's claim to be able to increase precipitation be seeding the clouds of suitable storms. Here the comparison is between seeding and not seeding. How can one obtain data to test the claim and to provide an estimate of the amount of increase?

As one possibility, the company might seed all suitable storms in the given location, and compare the resulting rainfall with that during the corresponding period in the preceding years. Of course, if the results are very striking (for example, if an enormous downpour occurs immediately following each seeding) this may settle the issue. However, typically the results are less clear. Suppose, for example, that the total rainfall matches, or even slightly exceeds, that of the wettest of the last five years. This may be the result of the seeding; or it may just be the consequence of an exceptionally wet year.

This difficulty is inherent in studies in which there is no randomness in the assignment to the experimental units of the conditions or treatments being compared. A better basis for the establishment of a causal relationship - in this case that the increased rainfall is due to the seeding - is obtained if storms are compared within the same season, and if they are assigned to the two treatments (seeding and not seeding) according to a random mechanism. This can be done in a variety of ways, corresponding to different experimental designs*.

As a simple possibility, suppose that the experiment is to extend to the first 20 storms that are suitable for seeding. Of these, 10 will be seeded and 10 not, the latter providing the controls. According to one design (complete randomization), ten of the numbers 1,...20 are selected at random; the storms bearing these numbers will be seeded, the remaining ten will not. An alternative design (paired comparisons*) pairs the storms (1,2), (3,4),...,(19,20), and within each pair assigns at random (e.g. by tossing a coin) one storm to seeding, the other to control. This second design will be particularly effective if storms occurring close together in time are more likely to be similar in strength than storms separated by longer time intervals.

It is interesting to note the close relationship of these designs to the sampling schemes of the preceding section. In the first design, the storms assigned to seeding are a simple random sample of the 20 available storms; in the second case, they constitute a stratified sample, with samples of size 1 from each of ten strata of size 2.

Unfortunately, random assignment of the conditions being compared is not always possible. Consider for example a study which reports that married men tend to live longer than unmarried ones. It is tempting to draw the conclusion that marriage prolongs life, perhaps by providing a more regulated life style. However, such a conclusion is not justified. Married and unmarried men constitute groups which differ in many ways. The latter for instance includes men with health problems which preclude marriage and which also tend to shorten life. It is thus not clear whether the observed effect is the result of the difference in marital status or of conditions leading to this difference. All that can be safely concluded is that marriage is associated with longer life expectancy. This may be enough for an insurance company but does not answer the sociological or public health question regarding the effect of marriage. Quite generally, observational studies* in which subjects have not been assigned to the conditions being compared (marital status, smoking habits, religious beliefs,...) can establish associations (such as that between marital status and longevity) but have great difficulty in validating causal relationships. To establish causation*+ is a cherished goal of statistical methodology but tends to be rather elusive. (For some further discussion of this point, see for example Mosteller and Tukey ([28], p.260/261).)

Example 6. A headache remedy. Suppose that you wake up with a headache, take a headache remedy, and an hour later find that the pain is gone. Based on this isolated instance (which in words attributed to R.A. Fisher is an experience rather than an experiment), it is clearly not possible to conclude that the result is due to the remedy. The cause might instead have been another intervening event such as breakfast, or possibly the headache had run its course and would have dissipated in any case.

The attribution of the cure to the medication becomes much stronger if the incident is not isolated. If you have suffered from headaches before, took the medicine sometimes soon after onset, at other times only after having waited in vain for the pain to subside on its own; if on different occasions you took the tablets at different times of day, and if under all these different circumstances the headache disappeared shortly after treatment and never (or only very rarely) before, the robustness of the effect over many different conditions would tend to carry conviction where a single instance would not. The finding could be strengthened further if your own experience could be merged with that of others. (However, even if the evidence for a treatment effect is convincing, observations of this kind can not determine whether the ingredients of the medication are responsible for the improvement, or whether it is a placebo effect, i.e. the patient's belief in the effectiveness of the treatment, which relieves the pain).

To see how to systematize the informal reasoning of the preceding paragraph, consider another example.

Example 7. New traffic signs. Suppose that 4-way stop signs have been installed at a dangerous intersection, and that four accidents had occurred in the month preceding the installment of the signs but only one in the month following it. These facts by themselves provide little basis for an inference. Suppose for example that the monthly accident frequencies constitute a purely random sequence of which the value 4 was a chance high, which however caused the installment of the signs. Then even without this intervention, a decrease could have been expected in the succeeding month.

A more meaningful comparison can be obtained if the accident statistics are available not only for one month before and after, but for several months in both directions. One is then dealing with an interrupted time series, for which it is possible to compare the observations before with those after the "interruption" (the installment of the signs). Even the nonrandom choice of the time of intervention would then have only a relatively minor effect.

While the interrupted time series can establish that the accident rate is lower after installment than it was before, it can not establish the new signs as the cause of the decrease since other changes might have occurred at the same time. To mention only one possibility, the community may have been affected by editorials published at the times of the third and fourth accidents. Some control - although not as firm as that resulting from randomization - can be obtained by studying the accident rates during the same months also at some other intersections, a multiple time series design. If the other series do not show a corresponding decease. this will clearly strengthen the causal argument

Quasi-experiments[1] such as the interrupted time series and multiple time series described above, which try to identify and control the most plausible alternatives to the treatment being the cause of the change, are treated by Cook and Campbell [9] and Cochran [8]; for an elementary discussion see Campbell [7]. They can greatly strengthen the causal attribution although they can not be expected to be as convincing as a controlled randomized experiment.

[1] A better term might be quasi-controlled experiments.

Serial data.

Most of the data considered so far were assumed to be collected on a one-shot basis. Often, they are obtained instead consecutively over a period of time. Such data are particularly useful in assessing changes over time. (Are winters getting colder? Is the birthrate in the U.S. falling? Are an author's novels becoming more popular?)

An important application is provided by statistical quality control*. An established production process is monitored by taking observations of the quality of the product at regular intervals. The process is said to be in control if the successive observations are independent and have the same distribution. A control chart* on which the successive observations are plotted provides signals when the process appears to be going out of control. A similar approach is used in monitoring the cardiogram of a heart patient in intensive care, where however the data consist of a continuous graph instead of a discrete sequence of points. Analogously, seismographs provide continuous data for the monitoring of seismic activity while persons watching their weight through regular weighings provide another illustration of the discrete case.

In these applications, a single process is followed to check that no changes are occurring and to alert us when they are. A different situation arises when one is interested in assessing the changes occurring in a population as part of a developing pattern or as the result of some intervention. For example, to study the pattern of growth (the growth curve*) of children, plants, or institutions, or the effect of several different programs (say, of different rehabilitation programs on juvenile offenders), one observes each member of a sample from the population over a period of time. In such longitudinal studies* the observations of the same units at different times often are dependent. Probability models for sequences of dependent observations are considerably more complicated than those for sequences i.i.d. variables. Such models and their statistical analysis are treated in the theory of time series* or more generally of stochastic processes*.

Observations arising serially, for instance on patients coming to a clinic, successive books by an authors, or stockmarket performance in successive time periods, provide an opportunity to economize by letting the size of the study be determined by the data. (according to a clearly specified rule), instead of fixing it in advance. Suppose for example that a shipment of goods is being sampled to determine whether its overall quality meets certain specifications. Items are being drawn at random, and examined, one by one. It may then be possible to stop early when the quality of the initial observations is either very satisfactory or very unsatisfactory, while one may wish to take a larger sample when the initial observations are mixed. The working out of economical stopping rules providing the desired statistical information, and the analysis of the resulting data, is the problem treated in sequential analysis*.

Designing experiments.

Part I of this article has been concerned with the interpretation of data once they have been obtained, Part II with various processes used to acquire the data. However, preceding even this stage, it is necessary to plan how many and what kind of data will be needed to answer the questions under consideration.

As an illustration, consider once more the ESP study of Example 2, based on 100 tosses of a coin. Suppose the investigator had decided, before the experiment was carried out, to give serious attention to the possibility of ESP provided the number of Subject A's correct calls would be at least 65. As was pointed out in Example 2, under the hypothesis of pure guessing the probability of 65 or more answers is .002. There is thus little danger of paying attention to the claim if it has no validity.

However, is this study giving Subject A a fair shake? What is the probability of getting 65 or more of the tosses right if A really does possess the claimed ability? The answer depends of course on the extent of the ability, which can be measured by the probability of calling a toss correctly. Here are some values, computed under the simplifying assumptions that the 100 calls are independent, and that the probability p of a correct call is the same for each toss.

p	.55	.6	.7	.75
$P(\text{No. correct calls} \geq 65)$ =Power	.027	.179	.884	.990

The probability in this table measures the power* of the test of the hypothesis $H: p = 1/2$, i.e. the probability of following up A's claim, against various alternative values of p. It shows that the power is quite satisfactory when p is .7 or more, but not, for example, when $p = .6$. In this case, despite the large discrepancy between pure guessing and an ability corresponding to $p = .6$, the probability is less than 20% that serious attention would be given to the claim. To get higher power would require a larger number of tosses. For example, to increase the power against $p = .6$ to .9, while keeping the probability of following up the claim when $p = .5$ at .002, would require a sample of about 425 tosses instead of the previous 100. A statement of the goals to be achieved, e.g. the required power of a test or accuracy of an estimate, and determination of the sample size needed for this purpose, are crucial aspects of planning a study.

Taking a fixed number of observations (100, or 425,...) is not necessarily the best design. Suppose for example that A calls all of the first 25 tosses correctly. This may be enough to provide the desired evidence of A's ability (or suggest that something is wrong with the experiment). Thus, a sequential stopping rule of the type indicated at the end of the last section might be more efficient.

Consider next the problem of determining an optimal design when different types of observations are involved. Suppose we are concerned with the comparison of two treatments, and that the total number of observations is fixed at N = 2k. Then it will typically be best to assign k subjects to each treatment, since this will tend to maximize the power of the tests and minimize the variance of the estimate of the difference. This may of course not be the case if observations on one of the treatments have smaller variance than on the other.

Finding the optimum assignment becomes much more difficult when observations are taken sequentially, and a decision is required at each stage which treatment to apply next. What is the best (or even a good) rule depends strongly on the purpose of the procedure. If the only concern is to decide whether the treatments are equally effective and, if not, which is better and by how much, the treatments should be assigned in a balanced way, i.e., so that each is received by the same number of subjects. However, if one is comparing two treatments of a serious medical condition, an additional consideration arises: a desire to minimize the number of patients in the study who receive the inferior treatment. A sequential procedure that has been suggested for such a case is the "play the winner rule" in which the next patient receives the treatment which at this point looks better. (For details, see for example, Flehinger and Louis [15] and Siegmund [32].) It should be noted however, that such an u require more observations than the best balanced rule, and hence will take longer to lead to termination of the study and hence to a recommendation. Thus, the inferior treatment will be assigned to fewer patients within the experimental group but to a larger number outside it.

Comparative experiments typically involve more than the study of just one difference. Consider an experiment to investigate the effect of a number of factors (to which experimental units can be assigned at random) on some observable response such as rainfall, crop yield, or length of life. Two possible designs are one-at-a-time experiments, in which each factor is studied in a separate experiment with all other factors held constant, and <u>factorial</u> <u>experiments</u>*, which provide a joint study of the effects of all factors simultaneously. The latter type of design has two important advantages.

Two or more effects may interact in the sense that the effect of one factor depends on the level of the other. In a study of the effect of textbooks and instructors on the performance of students, for example, it may turn out that a textbook which works very well for one instructor is quite uncongenial to another. The existence and size of such interactions* are most easily investigated by studying the various factors simultaneously. When no interactions are present, joint experimentation has the advantage of requiring far fewer observations to estimate or test the various effects than would be needed if each factor were studied separately. (For further discussion of these and related issues, see for example Cox [10] and Box, Hunter, and Hunter [5].)

A factorial experiment is concerned with the study of a number of factors, each of which can occur at several levels. A study of cancer treatments, for example, might involve different types of surgery, a number of surgeons, and several kinds of postoperative therapy (radiation, chemotherapy,...). The factorial experiment is said to be complete if one or more observations are obtained at all possible combinations of levels. If the number of such combinations is too large, the complete design may be replaced by a fractional factorial design*, in which only certain combinations of levels are observed. The theory of experimental design* is concerned with the search for good (or possibly optimal) designs able to provide the basis for a suitable analysis of the resulting data.

Unless the experimental units (the patients, students, agricultural plot,...) are fairly homogeneous, it may be advisable to divide them into more homogeneous strata, as was discussed earlier for the sampling of populations. In the present context, such stratification is called blocking*. The strata or blocks then play a role which in some respects is similar to that of an additional factor.

Much of the theory of factorial designs was originated by R.A. Fisher* in the 1920's, in the context of agricultural field trials, where it continues to play a central role; in addition, its uses have since expanded to many other areas of application.

Conclusion.

Statistics has been discussed in this article as dealing with the collection and interpretation of data to obtain information. We have been particularly concerned with the uncertainty caused by the fact that the observations might have taken on other values, and with the resulting variability of the data. An alternative view of statistics is obtained by noting that uncertainty attaches not only to data, but also to many other other 'chancy' events such as length of life, the occurrence of accidents, the quality of a manufactured article, or the fall of a die. Correspondingly, statistics has also been defined as the subject concerned with understanding, controlling, and reducing uncertainty. Actually, there is little difference between these two descriptions: Data involve uncertainty, and the study of any particular uncertainty requires the collection of appropriate data. It is a matter of emphasis.

Earlier sections have given a general indication of the pervasiveness and impact of statistical considerations relating to both data and uncertainty. To illustrate the power and wide range of the statistical approach we shall in this concluding section take a brief look at three specific studies.

Example 8. Polio vaccine trial. To test whether a proposed polio vaccine (the Salk vaccine) was effective, a large scale study was carried out in 1954. The most useful part of the study was a completely randomized trial in which about half of approximately 400,000 school-children were assigned at random to receive the vaccine, with the other half receiving a placebo (injection of an ineffective salt solution). The study was double blind, i.e. neither the children nor the physicians making the diagnosis knew to which of the two groups any given child belonged. The results in

	Total number	Number confirmed polio
Vaccinated	200,745	57
Placebo	201,229	142

Table 1. Adapted from Meier [25]

Table 1 show that vaccination has cut the rate of polio to about 40%. The probability of that marked a decrease under the hypothesis that the vaccine has no effect is about $1/10^7$, much too small to reasonably attribute the effect to pure chance. (Note that this example is of the same type as the hypothetical Example 2 (ESP).)

Example 9. Medical progress. In the preceding example we were concerned with the effectiveness of a single medical innovation - the Salk vaccine. The study reported in the present example (Gilbert, McPeek, and Mosteller [17]) addresses a much broader question: What can be said about the effectiveness of present-day medical (or rather more specifically, surgical and anesthetic) innovations as a whole? Such an investigation of the combined implications of a whole area of studies is called a meta-analysis. (For a recent account of Meta-analysis with many references to the literature, see Hedges and Olkin [20].)

A first step toward such an analysis is to decide what data to collect, in particular, which studies to include in the investigation. Ideally, one would like to have a list of all relevant studies for the period in question. For medical research, an approximation to this ideal is available in MEDLARS (Nat. Library of Medicine's MEDical Literature And Retrieval System), which provides exhaustive coverage of the world's medical literature since 1964. From MEDLARS, the authors obtained a set of 107 papers ("the sample") evaluating the success of specific innovative surgical and anesthetic treatments, and used these to assess the effectiveness of such treatments as a whole. The authors view this set of papers as a sample from the flow of such research studies, and therefore believe that the results should give a realistic idea of what to expect from future innovations, at least for the short term.

The sample contained a total of 48 comparisons of a new treatment with a standard (control). In 36 of the cases, the assignment to treatment and control was randomized, but not in the remaining 12. The results are summarized in the following table.

	Innovation highly preferred	Innovation preferred	About equal	Standard preferred	Standard highly preferred	Total
Randomized	5	7	14	6	4	36
Non-randomized	5	2	3	1	1	12

Table 3. Adapted from Gilbert, McPeek and Mosteller

A striking feature of the results for the randomized trials is that only 5 out of 36 or about 14% of the innovations were highly preferred. This suggests, the authors point out, that medical science is sufficiently well established so that substantial improvements over the standard treatments are difficult to achieve yet that it is not so settled that only a major theoretical advance will lead to any substantial further improvements.

A more sanguine assessment is obtained from the nonrandomized studies, where 5 out of 12 or about 42% of the innovations were highly preferred. Since randomization provides a surer foundation, the results in the first row may be deemed to be more reliable than those in the second. However, such a judgement overlooks the fact that the comparison of randomized with nonrandomized studies is itself not randomized, i.e. the assignment of studies to these two types was not, and was in fact far from, random. As a result, the two types of studies may not be directly comparable. For example, a surgeon who is strongly convinced of the great superiority of the new treatment, might for ethical or other reasons not be willing to apply the standard treatment, and would therefore have to look for controls from an earlier period or from other surgeons, while no such qualms might arise for a less dramatic innovation where a randomized study would then be acceptable. Many other explanations could be imagined, and much more information would be needed before one could attempt to assign a cause.

The authors go beyond the frequencies displayed in Table 3 to estimate the size of the treatment effects. For this purpose, they utilize a sophisticated technique, empirical Bayes*, which is particularly suited for meta-analysis.

Example 10. Literary detection. The Federalist papers are a historically important set of 85 short political essays published mostly anonymously in 1787/88 by Alexander Hamilton, James Madison, and John Jay. The authorship of most of the papers was eventually established but that of twelve of them has remained in dispute between Madison (M) and Hamilton (H), with historical research in all cases leaning towards, but not clearly deciding in favor of, Madison.

An effort to resolve the doubt by statistical methods was undertaken by Mosteller and Wallace [29]. The basic idea of such literary detection is to find aspects of the writing styles of the two (or more) authors in question which have good ability to discriminate between the various possibilities. This was particularly difficult in the present case because the two authors have very similar styles. However, by comparing known texts of M and H, Mosteller and Wallace were able to identify a number of words that were used with much higher frequency by one of the authors than the other. As an illustration Table 2 shows the frequency of occurrence of the word 'on' in blocks of about 200 words from texts by the two authors. (For example, the number of blocks in which the word 'on' occurred exactly twice was 27 for the 247 H-blocks, but 55 for the 262 M-blocks). The rate of occurrence of the word 'on' per thousand words of text was 3.38 for H and 7.57 for M.

	0	1	2	3	4	5	6	Total Number of blocks
H	145	67	27	7	1	-	-	247
M	63	80	55	32	20	8	4	262

Table 2 (Adapted from Mosteller and Wallace).

The present problem is similar to the diagnostic problem of Example 3, with θ taking on the two values H and M. (The statistical methodology dealing with the attribution of an item to one of two or more classes, a patient to a disease or a piece of writing to an author, is called classification^{*+} or discrimination.) Mosteller and Wallace's first task was to decide on observations X which might provide them with the evidence needed to settle the question. They selected a total of 30 words (including the word 'on') with high discriminating ability, and used as their observations X the frequencies with which these words occurred in a given text of disputed authorship.

Next the distribution $p_\theta(x)$ for the frequency of a given word had to be specified for each author with the help of the known H and M texts. Bayesian analyses for each of the 30 words based on the chosen family of distributions were combined into overall odds for M and H. For all but two of the papers, these overwhelmingly (several hundred million to one) favored Madison, given any reasonable prior odds for M and H. For each of these papers, the analysis leaves little room for doubt. In the remaining two cases Madison is also favored, but the odds are more modest, and the statistical attribution therefore less certain.

Statistical considerations arise in nearly all fields of human endeavor. Many of the associated activities are carried out by large numbers of full - or part - time professional statisticians. Some of the demands and satisfactions of a statistical career are described by Healy [19], in the pamphlet: Careers in Statistics (Revised 3-rd ed., 1980) published by the American Statistical Association, and in Deming [11]. More detailed descriptions of the work and requirements of statisticians at various levels in Government and Industry can be found in articles on Preparing Statisticians for Careers in Federal Government and in Industry (Amer. Statistician, vol. 36 (1982) 69-89 and vol. 34 (1980) 65-80, and in Moser [26].

The basic training in statistics, as in most other fields, occurs at Universities and Colleges where Departments of Statistics offer both undergraduate and graduate degrees in statistics. (At some institutions, the principal statistics courses are instead provided by the mathematics department.) In addition, there may be degrees in biostatistics*. Statistics courses and quantitative programs with a strong statistical component may also be offered in Operations Research, Business Schools, Demography, Economics, Education, Psychology, and Sociology. In addition to courses and programs preparing for a profession in statistics, Statistics Departments also provide 'service courses' both at introductory and advanced levels to students in other fields who may need to use statistical methods in their work.

Another need of statistical education is filled by courses in statistical concepts as part of a general education. Acquiring the ability to think in statistical terms is of great importance even for persons without a quantitative bent in view of the pervasive occurrence of statistical ideas in newspapers and magazines, and in the terminology we use to describe and discuss the world around us. What do we mean by saying that women tend to live longer than men, or that cancer patients survive longer today than ten years ago? Does it mean they live longer on the average, that the median length of their life is longer, or that they have a better chance to survive to any given age? And is the longer survival of persons diagnosed to have cancer primarily due to the availability of more effective treatments, including earlier diagnosis? In fact, is there even an increase in length of survival, or is the apparent increase just a statistical consequence of earlier diagnosis? If the disease is diagnosed a year earlier, the survival after diagnosis has increased by a year even if nothing else has changed.

To consider another example, what is meant by the rate of unemployment? Roughly speaking, it is the proportion among the people wishing to be employed, who are not. But do the official rates include those past seekers for jobs who have given up in despair? A change in the figure for unemployment may be the result of a small change in the definition, or of some sociological change such as the increased entry of women into the work force.

Many of the considerations involved in such issues are statistical. Fortunately, the basic ideas needed for general discussion and comprehension can be communicated at a fairly nontechnical level, as is done, for example, in Tanur et al [35], Mosteller et al [27], and Freedman, Pisani, and Purves [16]. Books such as these, and the increasing availability of first courses in probability and statistics in High Schools, should have the effect of gradually raising the general level of statistical literacy.

Acknowledgment. I should like to thank David Cox, Persi Diaconis, David Freedman, William Kruskal, Frederick Mosteller, Juliet Shaffer, and the editors for reading a draft of this article, and providing me with comments and

suggestions which resulted in many improvements.

REFERENCES

[1] Alterman, H. (1969). Counting People. Harcourt, Brace and World. New York.

[2] Barnett, V. (1982). Comparative Statistical Inference. 2nd Ed. John Wiley, N.Y.

[3] Berger, J. (1984). In: Robustness of Bayesian Analysis. (Kadane, Ed.) North Holland. Amsterdam.

[4] Berger, J. (1985). Statistical Decision Theory and Bayesian Analysis. 2nd Ed. Springer, New York.

[5] Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). Statistics for Experimenters, John Wiley, New York.

[6] Campbell, W.E. (1942). Bell Telephone System, Technical Public. Monograph B-1350.

[7] Campbell, D.T. (1978). In: Statistics: A Guide to the Unknown (Tanur et al, Eds), Wadsworth, Belmont.

[8] Cochran, W.G. (1983). Planning and Analysis of Observational Studies, Wiley, New York.

[9] Cook, T.D. and Campbell, D.T. (1979). Quasi-Experimentation. Rand McNally, Chicago.

[10] Cox, D.R. (1958). Planning of Experiments. John Wiley, New York.

[11] Deming, W.E. (1986). In Encyclopedia of Statistical Sciences, Vol. 7.

[12] Diaconis, P. (1985). Theories of data analysis, In [23]

[13] Dickey, J.M. (1973). *J. Roy. Stat. Soc. (B),* **35**, 285-305.

[14] Feynman, R. (1965). The Character of Physical Law. The British Broadcasting Corp.

[15] Flehinger, B.J. and Louis, T.A. (1972). *Biometrika,* **58**, 419-426.

[16] Freedman, D., Pisani, R. and Purves, R. (1978). Statistics, Norton, New York.

[17] Gilbert, J.P., McPeek, B. and Mosteller, F. (1977). Progress in surgery and anethesia. In: Costs, Risks and Benefits of Surgery (Bunker et al, Eds), Oxford Univ. Press. New York.

[18] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). Robust Statistics, Wiley, New York.

[19] Healy, M.J.R. (1973). *J. Roy. Statist. Soc. (A),* **136**, 71-74.

[20] Hedges, L.V. and Olkin, I. (1985). Statistical Methods for Meta-Analysis, Academic Press, Orlando.

[21] Hoaglin, D.C., Light, R., McPeek, B., Mosteller, F. and Stoto, M.A. (1982). Data for Decisions, Abt, Assoc. Cambridge.

[22] Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (1983). Understanding Robust and Exploratory Data Analysis, Wiley, New York.

[23] Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (1985). Exploring Data Tables, Trends, and Shapes, Wiley, New York.

[24] Huff, D. (1954). How to lie with statistics, Norton, New York.

[25] Meier, P. (1978). In: Statistics: Guide to the Unknown. 2nd Ed. (Tanur et al, Eds).

[26] Moser, C.A. (1973). *J. Roy. Statist. Soc. (A),* **136**, 75-88.

[27] Mosteller, F., Kruskal, W.H., Link, R.F., Pieters, R.S. and Rising, G.R., Eds (1973). Statistics by Example (4 vols), Addison-Wesley, Reading.

[28] Mosteller, F. and Tukey, J.W. (1977). Data Analysis and Regression, Addison-Wesley, Reading.

[29] Mosteller, F. and Wallace, D.L. (1984). Applied Bayesian and Classical Inference, Springer, New York.

[30] Rubin, D.B. (1984). *Ann. Statist.,* **12**, 1151-1172.

[31] Shortliffe, E.H. (1976). Computer-based Medical Consultations, MYCIN, Elsevier, N.Y.

[32] Siegmund, D. (1985). Sequential Analysis, Springer, New York.

[33] Smith, A.F.M. (1984). *J. Roy. Statist. (A),* **147**, 245-259.

[34] Spiegelhalter, D.J. and Knill-Jones, R.P. (1984). *J. Roy. Stat. Soc. (A)*, **147**, 35-77.

[35] Tanur, J., Mosteller, F., Kruskal, W.H., Link, R.F., Pieters, R.S., Rising, G.R. and Lehmann, E.L., Eds (1978). Statistics: A Guide to the Unknown (2nd Ed), Wadsworth, Belmont.

[36] Wolfowitz, J. (1943). *Ann. Math. Statist.*, **14**, 280-288.