Statistics: Challenges and Opportunities for the Twenty-First Century

Edited by: Jon Kettenring, Bruce Lindsay, & David Siegmund

Draft: 6 April 2003

Contents

An executive summary

On May 6-8, 2002 approximately fifty statisticians from around the world gathered at the National Science Foundation to identify the future challenges and opportunities for the statistics profession. The workshop was largely focused on scientific research, but all participants were asked to discuss the role of education and training in attaining our long term goals as a profession. The scientific committee was placed in charge of producing a workshop report.

A substantial proportion of the report was devoted to describing the unique role of statistics as a tool in gaining knowledge, with the goal of making the report more accessible to the wider audience of its key stakeholders, including universities and funding agencies. This was done largely because the role of statistical science is often poorly understood by the rest of the scientific community. Much of the intellectual excitement of the core of the subject comes from the development and use of sophisticated mathematical and computational tools, and so falls beyond the ken of all but a few scientists.

For this reason there is a potential confusion about how statistics relates to mathematics, and so a portion of the report dealt with separating the identity of the two. Statistics is no longer, if it ever was, just another mathematical area like topology, but rather it is a large scale user of mathematical and computational tools with a focused scientific agenda.

The report goes on to identify key opportunities and needs facing the statistics profession over the next few years. In many ways, the main issues arise from the tremendous success of statistics in modern science and technology. Growth is occurring in virtually every direction. We face increased number of students, increased opportunities for collaboration, increased size of data sets, and the need for expanded sets of tools from mathematics, computer science, and subject matter areas.

An important aspect of this growth has been its tremendous impact on the breadth of our research activity. We now face a multitude of exciting research possibilities for the next century. This report give examples of many of these topics, both in the core research areas of statistics and in the many disciplines where statistical research is carried out, such as biology, engineering and industrial science, geological and environmental science, information technology, physical sciences, and social and economic statistics.

This growth and success has also created stresses on the profession. As a result, the workshop report identifies the following areas where the pro-

fession and its stakeholders should put future effort:

- **Promoting the unique identity of statistics**. This involves clarifying its role *vis à vis* the other sciences, especially in clarifying the differences in roles and needs between statistics and the other mathematical areas. Doing so will make it easier for the profession to access the tools needed to handle growth.

- **Strengthening the core research areas**. This requires ensuring that their full value in consolidation of knowledge is recognized and rewarded. Achieving this goal would be aided by expansion of the current funding programs in this area.

- **Strengthening multidisciplinary research activities**. This can be accomplished by helping all parties involved calculate the full cost of multidisciplinary research and so make allowances for its special needs. This would imply research support that recognizes the special role of statisticians in scientific applications, where they are toolmakers (concepts, models, and software) as well as tool providers.

- **Developing new models for statistical education**. This should be done at every level, and separately for majors and non-majors. Attaining this goal will require the focused efforts of experts in statistical education working together to create and disseminate the models.

- **Accelerating the recruitment of the next generation**. For this to succeed, the profession needs the cooperation of the key stakeholders in developing programs that recognize the special needs and opportunities for statistics.

The many issues and steps involved in attaining these goals are described in more detail in the report.

1

Introduction

1.1 The workshop

A workshop on the future of statistics was held at the National Science Foundation in early May of 2002. This report presents the general conclusions that arose at that time, along with supplemental supporting material. The workshop was held at the request of the Foundation, and was organized by a scientific committee of 9 members. That same committee has prepared this report with the guidance and assistance of the workshop participants.

There were about fifty participants in the workshop, chosen to represent the breadth of the statistical profession. There were a significant number of non U.S. participants, but on the whole the workshop and the report are focused on the statistical sciences within the boundaries of the United States. The names of the participants, together with the names of the scientific committee, are included within Appendix A to this report. This appendix also contains the schedule of talks at the workshop, as well as the charge delivered to the participants.

The workshop took place at a time of great ferment and excitement in the world of statistics. Recognizing this, the National Science Foundation supported the workshop as a means for the community to come together to identify its common needs, goals, and aspirations. The scientific committee was given a free hand to create the workshop as well as to design the final product, this report. However, it was clear that the sponsors sought a product that would have a wide and positive effect on the future of statistics in the United States.

The committee has decided that, for maximum impact, this report should be directed to a wide range of audiences. Thus our target is not just statistics researchers, but also such important supporting players as collaborators, department heads, college deans, and funding agencies. The workshop participants, themselves from a diverse set of scientific frontiers, found much in common about the challenges and opportunities facing their profession. This report is designed to encompass this common ground in a manner intelligible to the whole scientific community.

There is one important area of statistics that fell outside the charge of the workshop. Biostatistics, which is statistics as applied in the health professions, is a large and thriving subdiscipline. Although it shares with the mainstream of statistics both philosophy and tools, the funding and focus of biostatistics departments are sufficiently separate from the rest of statistics to not try to put them under the same umbrella of needs and

issues. On the other hand, many of the issues presented in this report should be of high relevance to this area as well.

1.2 What is statistics?

Given that we were addressing a wide audience, the committee felt some need to clarify the role of statistics in science. Many scientists have seen our profession only peripherally,if at all. To fulfill this need, the first speaker at the workshop, the eminent Professor D. R. Cox of Oxford University, was asked to start with the basics and identify "What is Statistics?". This question was to be repeatedly addressed through the course of the workshop. We summarize some of the key points here.

The subject of statistics is distinguished by its essentially multidisciplinary nature. The over-arching goal of the profession is the extraction of scientifically meaningful information from data of all types. Statisticians carry out this goal in a variety of ways and at a number of stages of the scientific process. Being trained in principles of good experimentation, they can be collaborators in the scientific process from the beginning experimental design all the way through to the final data analysis and conclusions.

The scientific domains of this work are nearly as wide as all scientific endeavor. In this report we focus on six main areas: the core of statistics plus six principal areas of application:

- biological science
- engineering and industrial statistics
- geological and environmental sciences
- information technology
- physical sciences
- social and economic science

These categories were chosen to correspond roughly to the different directorates of the National Science Foundation in which the research is supported. A seventh large area, health-related statistics, usually called "biostatistics", derives much of its funding from the National Institutes of Health. Although much of the content in this report is relevant to this area, there is enough unique to biostatistics to make it deserving of a separate report.

Outside the collaborative domain, the core activity of statisticians is the construction of the mathematical and conceptual tools that can be used for information extraction. Much of the research has as its mathematical

basis probability theory, but the end goal is always to provide results useful in empirical work. This distinguishes the theoretical research efforts of statisticians from most areas of mathematics in which abstract results are pursued purely for their intrinsic significance. As was stated in the NSF Report 98-95, "Report of the Senior Assessment Panel for the International Assessment of the U.S. Mathematical Sciences", hereafter called the "Odom Report,":

> Statistics has always been tied to applications, and the significance of results, even in theoretical statistics, is strongly dependent on the class of applications to which the results are relevant. In this aspect it **strongly differs** from all other disciplines of the mathematical sciences except computational mathematics. (Our emphasis)

A distinguishing feature of the statistics profession, and the methodology it develops, is the focus on logic-based principles for drawing scientific conclusions from data. This principled approach distinguishes statistics from a larger venue of data manipulation, organization, and analysis. One key principle of the profession dictates that one should proceed with measurable caution in the search for scientific truths within data. Such statistical tools as confidence coefficients, significance levels, and credible regions were designed to provide easily interpreted measures of validity. When used appropriately, these tools help to curb false conclusions from data.

Benjamin Disraeli, later quoted by Mark Twain, said, "There are three kinds of lies: lies, damned lies, and statistics." In fact, statisticians are trained to operate at the other end of the spectrum, separating scientific truth from scientific fiction. To illustrate this point, later in this report we will discuss a new measure of validity, the *false discovery rate*, that was developed due to the massive data sets and wide range of hypotheses that occur in modern scientific investigations.

Of course, statisticians do not own the tools of statistics any more than mathematicians own mathematics. Certainly most statistical applications and much statistical research is carried out by scientists in other subject matter areas. The essential role of statistical research is to develop new tools for use at the frontiers of science. In the later sections of this report we will demonstrate the very exciting statistical research possibilities that have arisen in recent years. In particular, the possibilities for data collection and storage have opened the need for whole new approaches to data analysis problems.

1.3 The statistical community

By the nature of their work, statisticians work in a wide array of environments. In the United States there are many statisticians who work in

Departments of Statistics. Such departments are found at most of the major research universities. There are now 86 Ph.D. programs in Statistics, Biostatistics, and Biometrics. They have tended to focus on graduate research, including collaboration with other disciplines, and education, as well as undergraduate service courses. One key question to be addressed later in this report is their potential future role in providing undergraduate majors in statistics as a part of a major effort to increase the size of the pipeline into the profession.

These departments largely arose by splitting off from mathematics departments in the second half of the twentieth century. As such statistics is often viewed as a branch of mathematics. This structural view is evidenced in the National Science Foundation itself, in which Probability and Statistics is one branch of the Division of Mathematical Sciences, placed side by side with such "pure" branches as Topology and Algebra. However one of the key conclusions of the participants of the Futures workshop was that statistics has become more and more distinct from the other mathematical areas. The scientific goals of statisticians and the directions of modern science point to a world where computer and information science tools are at least as important to statistics as those of probability theory.

A substantial fraction of the academic statistics community works in departments other than statistics. This can occur even in universities with statistics departments, where they can be found in business schools, social science and science departments across the spectrum. In schools without statistics departments, as for example in four year colleges, there are often statisticians within the mathematics department where they are needed for undergraduate education. Finally, there are also many statisticians who work in biostatistics departments.

Going beyond the academic community, but well connected to it, are many more statisticians employed in government and business, as well as many users of statistics. The NSF Report 98-95, the Odom Report, stated regarding the field of statistics:

> The interaction between the academic community and users in industry and government is highly developed, and hence there is a rapid dissemination of theoretical ideas and of challenging problems from applications, as well as a tradition of interdisciplinary work.

Statisticians are found in government agencies from the Census Bureau to the National Institute of Standards and Technology to the National Institutes of Health. They are employed across a wide range of industries, often for quality control work. In particular, the pharmaceutical industry has been a leading employer of statisticians, who carry out the design and analysis of experiments required for drug development.

1.4 Resources

This futures report appears to be the first of its kind for statistics, and so we cannot provide benchmarks from past reports for comparison. Our primary sources of information for the health of the profession, as well as its trends are materials from the joint mathematical societies and from the National Science Foundation. Within these documents it is important to separate out the information that is particular to statistics from that of mathematical sciences as a whole, as in many cases the trends are different.

The NSF Report 98-95 titled "Report of the Senior Assessment Panel for the International Assessment of the U.S. Mathematical Sciences," widely known as the Odom Report, is an important document because of its role in generating policy at the Foundation. It provided an independent assessment of the needs of the mathematical sciences including statistics, but for the most part focuses on mathematics as a whole. There are sufficient commonalities between the needs of mathematics and statistics to use it as an important source document. For example, the report identifies the three primary activities of mathematicians as being:

1. Generating concepts in fundamental mathematics
2. Interacting with areas that use mathematics, such as science, engineering, technology, finance, and national security; and
3. Attracting and developing the next generation of mathematicians.

After substituting statistics for mathematics, this trichotomy serves well to describe the primary activities of statisticians as well. One fundamental distinction between mathematics and statistics lies in the balance between items 1 and 2, as will be discussed later.

2

Historical Overview

Statistical methods have a long history of application in the sciences, although its recognition as a separate field dates largely from the twentieth century. Stigler (1986) identifies modern statistics as unified subject, "both a logic and a methodology", that grew out of a diversity of ideas. One stream of the story is data analytic, as it arises from the problem of combining measurements in astronomy and geodesy. Among the earliest contributions was the development of the method of least squares by Legendre around 1800.

A second stream, the basis for the theory of uncertainty, arose from the early developments in the theory of probability. Here the mathematicians Bernoulli, DeMoivre, Bayes, Laplace, and finally Gauss laid the foundations for the construction of probability models, as well as provided a basis for inverting probability models to draw conclusions about data.

The late part of the nineteenth century brought a fundamental coalescence of statistical thinking in England, but now the measurements that generated the concepts were those of heredity and biometrics. The key statistical ideas of correlation and regression were developed at this time. Soon thereafter, the chi-squared test was developed by Karl Pearson (1900). This was a tremendously important conceptual breakthrough; it is still being used for the rigorous testing of scientific hypotheses within a statistical model. The Department of Applied Statistics at University College in London was founded in 1911 by Karl Pearson, and was the first university statistics department in the world. It arose from the merger of a Eugenics Laboratory and a Biometric Laboratory.

Within a few years, R. A. Fisher, also of England, created the foundations of much of modern statistics. Fisher, also the founder of modern population genetics, was a genius of the highest order. He established methods for the analysis of complex experiments, now called "analysis of variance", which are used thousands of times each day by scientists around the globe. He showed that a function he called the likelihood could be used to develop optimal estimation and testing procedures in almost any probability model. He founded and developed the main ideas in the design of experiments, inspired by his work on agricultural field trials.

Fisher had a tremendous statistical intuition. At least some of the important work of the twentieth century was simply the attempt to clarify the significance and expand the domain of his groundbreaking research. Among the important works that followed in the 1930's was the rigorous development of the theory of hypothesis testing by Jerzy Neyman and Egon

Pearson at University College London. This theory became the foundation of research in this area for the remainder of the twentieth century.

By the mid-century, statisticians in the United States were making seminal contributions. Abraham Wald of Columbia University was a leader in the formal development of sequential analysis, a subject that grew out of the need for efficient sampling methods during World War II. Wald was also a leader in the development of decision-theoretic methods in Statistics. Another important player of this period was C. R. Rao, now of Pennsylvania State University and recent winner of the National Medal of Science, who produced many innovations in multivariate analysis, which is the study of the complex structures that exist in data of high dimensions. Another Medal of Science winner, John Tukey of Princeton, is the father of modern data analysis.

It was also during this period that statistics started to become institutionalized as a separate subject in the United States, distinct from the rest of mathematics or particular areas of application. In the U.S., Columbia (1946) and University of North Carolina (1946) were among the earliest departments. Through the rest of the century, as the role of statistics in the sciences expanded, the number and size of such departments grew. In the next section we will provide indicators for this growth.

Many important advances of the past century came in the area of modelling and estimation, where methods were developed that expanded the horizon of possible models and widened the range of validity of statistical procedures. An important adjunct to these developments was the wide expansion of so-called large sample theory, the study of the distributional properties of statistical procedures when the sample sizes are large. Accurate measures of uncertainty are the key components of statistical inference; large sample methods have enabled statisticians to calculate excellent approximations to these measures in a very wide range of problems.

Beginning in the 1970's, a major revolution in science occurred; it was destined to change the face of statistics forever. Beginning with punch cards and awkwardness, but rapidly replacing the existing slower alternatives, the computer has changed completely what it means to carry out a statistical analysis. It has also changed the facility with which one can collect and store data.

What are the consequences of this? This report is about those issues. We conclude our history by noting that the most successful methodologies at the end of the twentieth century, such as the bootstrap, and the proportional hazards model, would have been impractical without these changes in computing power. And the capacity for scientists to collect ever more data as well as data of greater sophistication points to an exciting and challenging future with more fundamental results.

3

Current Status

3.1 General overview

The theory and application of statistics spreads across many disciplines. The workshop organizers asked the following eminent speakers to deliver keynote talks on the following subareas:

- Core of statistics: Iain Johnstone, Stanford University.
- Biological statistics: Warren Ewens, University of Pennsylvania.
- Engineering and industrial statistics: Vijay Nair, University of Michigan.
- Geological and environmental statistics: Richard Smith, University of North Carolina.
- Information technology and statistics: Werner Stuetzle, University of Washington.
- Social and economic statistics: Joel Horowitz, Northwestern University.

In addition, there were talks by Chris Heyde, Columbia University and James Berger, Duke University on "Statistics in the international scene" and "Institutes: the role and contribution to statistics" respectively.

It should be noted that the subject of statistics does not have an agreed upon division of its heritage into distinct areas of research, such as "algebra, analysis, topology, and geometry" are in mathematics or "inorganic, organic, physical, and biophysical" are in chemistry. There is rather a central portion of the research, that we will call the core, and applications-oriented research that we have divided by scientific field.

The workshop organizers took these lectures and the ensuing discussions as the basis for the following report on the current status of the statistics profession. We start with some general statements about the state of the profession, and then turn to more specialized issues organized by the above research areas.

3.1.1 The quality of the profession

The Odom Report provided a strong endorsement to the quality of the U.S. effort in statistics, stating that: "the statistical sciences are very healthy

across all subareas in the United States, which is the clear world leader."

An informal survey of four leading statistics journals (two of which are based in the United Kingdom) substantiates this statement. The following table shows the departmental affiliation of the the U.S. based authors in these journals.

Statistics	**49%**
Biostatistics	**23%**
Industry	**6**%
Math.Science	**5**%
Mathematics	**4**%
Other	**13**%

Approximately one-half of the the authors had U.S. affiliations. Essentially all of these authors are in academic institutions. Moreover, the vast majority come from statistics or biostatistics departments, with less than one in ten coming from a department of Mathematics or Mathematical Sciences. The next table shows the reported sources of funding for this published research:

NIH	**40**%
NSF	**38**%
NSA	**9**%
ARO/ONR/EPA	**4**%
Other	**9**%

Clearly the National Science Foundation and the National Institutes of Health are the major role players in funding research in statistics. .

3.1.2 The size of the profession

One way to gauge the size of the statistics profession is to compare it with the rest of mathematics. In the following table we give the approximate number of members in the leading statistics and mathematics societies.

American Statistical Association (ASA)	**16,000**
Institute of Mathematical Statistics (IMS)	**3,500**
Biometric Society (ENAR/WNAR)	**3,500**
American Mathematical Society (AMS)	**30,000**
Mathematical Association of America (MAA)	**33,000**
Society for Industrial and Applied Math (SIAM)	**9,000**

These numbers are somewhat difficult to compare due to overlapping membership lists, but they do suggest that the number of statistics professionals might be somewhere between one-fourth to one-half the number of mathematicians.

The American Mathematical Society annual survey of 2001 indicates that there are 86 doctoral programs in statistics, biostatistics, and biometrics

(Group IV). This can be compared with 196 programs in other areas of mathematics (Groups I, II, III, V). Again, the numbers are not easy to compare, but do provide some idea of the scale.

A better measure might be the annual number of Statistics Ph.D's. However, these counts suffer from many of the usual data collection challenges: definition of population, quality of data, and census non-response. The following table presents three rather different numbers for statistics, as well as two estimates for the rest of mathematics.

AMS Survey 2000 (excluding probability)	**310**
Amstat Online 2000 (self reports)	**457**
NSF Survey of Earned Doctorates 2000 (accumulated over statistical subfields)	**822**
For reference, math excluding statistics:	
AMS Survey 2000	**809**
NSF Survey of Earned Doctorates	**925**

The AMS survey acknowledges problems with non-response from Statistics programs. The NSF Survey of Earned Doctorates number is derived by aggregating "statistical subfields" from the nearly 300 fine categories by which fields of doctorate are categorized in this Foundation wide survey.

If we consider the number of doctorates in math excluding statistics, there is greater coherence between the AMS and NSF surveys, again suggesting problems with the identification and collection of data for statistics in particular.

The NSF survey does provide data back in time that is useful for understanding how the relationship between statistics and the rest of mathematics has changed over the last 35 years. Figure 3.1 shows that the annual number of statistics Ph.D.s (per NSF definition) started at 200, less than 1/3 the number of mathematics degrees, but has grown more or less linearly ever since to 800, staying roughly equal with mathematics in the 1980's, and surpassing mathematics slightly since.

The number of research doctorates is a noisy surrogate for the level of research activity. Regardless, there are three program directors in DMS for statistics and probability as opposed to 19 for all other mathematical areas. This balance does not seem to reflect the magnitude of the statistics research effort, as measured by participants, nor its importance to science, as we will demonstrate.

3.1.3 The Odom Report: Issues in mathematics and statistics

The Odom report provided some broad statements about the most important issues in mathematics as a whole. In this section we discuss them in the context of the current status of statistics. We will later revisit these themes in the designated subareas.

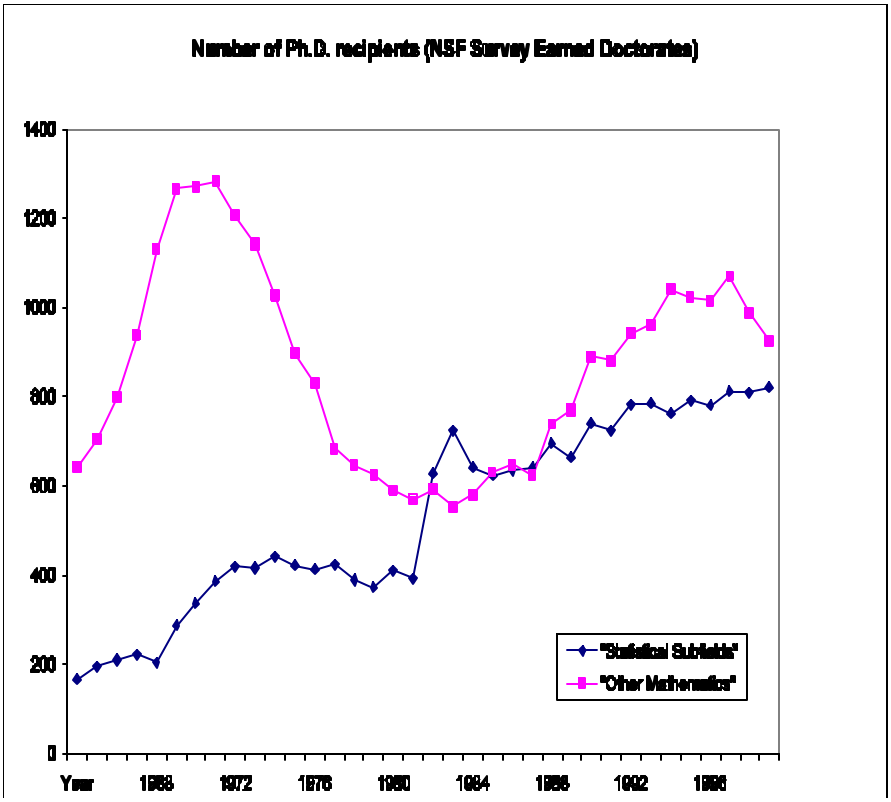

FIGURE 3.1. NSF Survey on the Number of Doctorates by Subject Matter

Data collection

A major theme of our report is that the statistics profession is experiencing a dramatic growth in its scientific value and its scientific workload due to changes in science, and in particular data collection. The Odom Report stated that

> With the advent of high-speed computers and sensors, some experimental sciences can now generate enormous volumes of data–the human genome is an example–and the new tools needed to organize this data and extract significant information from it will depend on the mathematical sciences.

Of all the mathematical sciences, the statistical sciences are uniquely focused on the collection and analysis of scientific data. Every senior statistician has felt the impact of this startling growth in the scale of data in recent years.

Increased opportunities for scientific collaboration

A second major theme of this report is that concurrent with the increased demand for statistical knowledge in the sciences comes an increased pressure for statisticians to make major time commitments to gaining knowledge and providing expertise in a diverse set of scientific areas. As noted in the Odom Report,

> Both in applications and in multidisciplinary projects ... there exist serious problems in the misuse of statistical models and in the quality of the education of scientists, engineers, social scientists, and other users of statistical methods. As observations generate more data, it will be essential to resolve this problem, perhaps by routinely including statisticians on research teams.

The Odom report further noted the scientific problems of the future will be extremely complex, and require collaborative efforts. It states that it will be virtually impossible for a single researcher to maintain sufficient expertise in both mathematics/computer science and a scientific discipline to model complex problems alone. We wholeheartedly agree with this finding, and will elaborate on it further.

The next generation

In several ways the future challenges to statistics differ from that of mathematics. For example, the Odom Report identifies three key issues:

> ...the mathematics community in the United States shares with other nations significant disciplinary challenges including a condition of isolation from other fields of science and engineering, **a decline in the number of young people entering the field**, and a low level of interaction with nonacademic fields, especially in the private sector. (emphasis ours)

It is certainly our observation that the number of U.S. residents entering the statistics field has shrunken over the years, and the growth in Ph.D. degrees has come largely by foreign recruitment. On the other hand, in the opinion of the scientific committee, the Odom Report's concern about isolation from other fields, scientific and non-scientific, seems not to apply to the current statistics scene.

4

The Core of Statistics

Statistics has an expanding intellectual heritage that we might, for lack of a better word, call the *core of statistics*. This terminology is not routinely used in the profession and so it is necessary to define rather precisely what is meant. We will define the core of statistics as the subset of statistical activity that is focused inward, on the subject itself, rather than outward, towards the needs of statistics in particular scientific domains. As a synonym for "core" the word "inreach" might be offered. This would reflect the fact that this core activity is the opposite of outreach. As such, almost all statisticians are active in both inreach and outreach activities.

The research in the core area is focused on the development of statistical models, methods, and related theory based on the general principles of the field. The objectives are to create unifying philosophies, concepts, statistical methods, and computational tools. Although this is introspective activity, a central philosophy of the core is that the importance of a problem is not dictated by its intrinsic beauty (as, say, in abstract mathematics). Rather, its importance is dictated by its potential for wide application or, alternatively, for its value in expanding understanding of the scientific validity of our methods.

Through this combination of looking inward and looking outward, the core serves very much as an information hub. It is defined by its connectivity to, and simultaneous use in, virtually all other sciences. That core statistical concepts and methodology can be used simultaneously in a vast range of sciences and applications is a great source of efficiency in statistics, and as a consequence, provides high value to all of science.

Core research might be contrasted to "application-specific statistical research", which is more closely driven by the need to analyze data so as to answer questions in a particular scientific field. Of necessity, this research draws on core knowledge for tools as well as for an understanding of the limitations of the tools. It also provides raw material for future core research through its unmet needs.

4.1 Understanding core interactivity

One way to demonstrate the amazing way that the core activities of statistics provide widespread value to the scientific community is to consider data on the citations of statistical literature. We offer an initial caution that citation data should not be overinterpreted, as high citations for individual

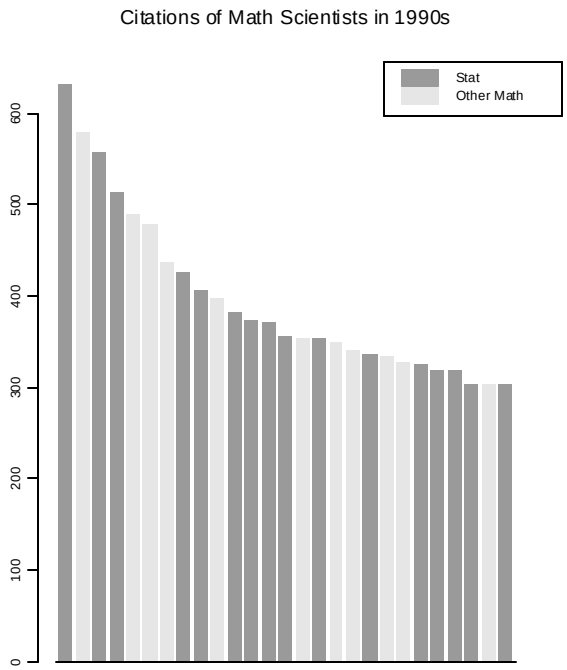

FIGURE 4.1. Citation counts of the most cited mathematical scientists

articles can reflect things other than quality or intrinsic importance. We offer citation data here because it provides a simple and accessible measure of the widespread influence of statistical research on scientific fields outside of statistics.

The Institute of Scientific Information (ISI), which produces the Science Citation Index and its relatives, created several lists of the "most cited scientists in the 1990's." Their data on mathematicians, published in the May/June 2002 issue of Science Watch, showed that eighteen of the twenty-five most cited mathematical scientists of the period 1991 to 2001 were statisticians or biostatisticians. Citation counts per author are given in Figure 4.1 In addition, the Journal of the American Statistical Association was far and away the most cited mathematical science journal.

There is evidence that this high rate of citation of statistical articles, relative to mathematics as a whole, is related to its wide scientific influence. For example, the paper by Hall and Titterington (1987), which considers the thorny problem of choosing a smoothing parameter in nonparametric function estimation, has about 2/3 of its citations outside any definition of

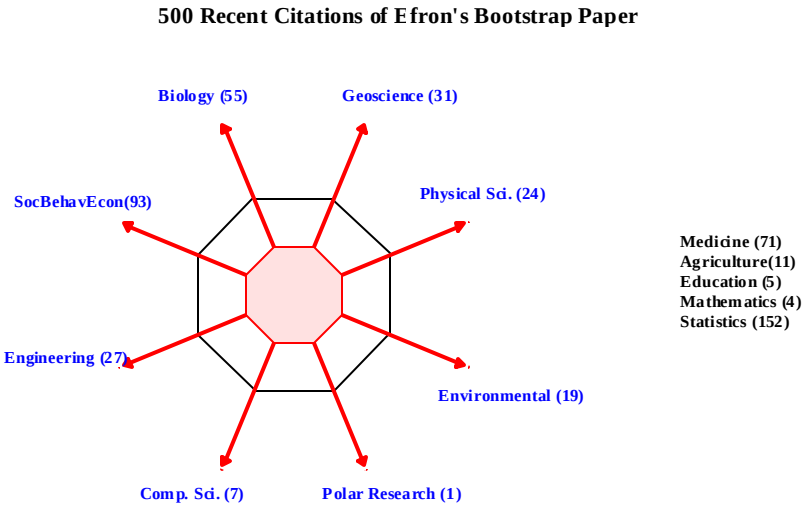

FIGURE 4.2. The dispersal of statistical information to other disciplines

the core of statistics, including the IEEE journals, J. Microscopy, Biomedical Engineering, and Journal de Physique. This is despite its appearance in a core research journal, and its theoretical cast.

One of the most important articles that leapt directly from core research into the mainstream of many scientific areas is the one introducing bootstrap methods. An examination of 500 recent citations of this paper shows that only 152 of these citations appeared in the statistics literature. Figure 4.2 shows the wide dispersal of this innovation that was generated in the core of statistics.

Of course, the core also arises at meaningful and useful methods for science because it reaches out to specific areas, finds important ideas, and creates the necessary generalizations that widen applicability. As an example, we might consider the development of methods that had their origins in age-specific death rates in actuarial work. In 1972 and 1975 the ideas of proportional hazards regression and partial likelihood analyses were introduced, which greatly enriched the tools available for the analysis of lifetime data when one has censored data along with covariate information. Since that time, these ideas and this methodology have grown and spread throughout the sciences to all settings where data that is censored or partially observed occurs. This would include astronomy, for example, where a star visible with one measurement tool might be invisible due to inadequate signal with a second measurement tool.

4.2 A detailed example of interplay

The following recent example illustrates in more detail the theme that the core research in statistics feeds off and interacts with outreach efforts. Since at least some of the work is NSF funded, it indicates in part the kind of interactions that should be kept in mind when supporting core research.

Last year, three astrophysicists published in *Science* a confirmation of the Big Bang theory of the creation of the universe. They studied the imprint of so-called acoustic oscillations on the distribution of matter in the universe today and showed it was in concordance with the distribution of cosmic microwave background radiation from the early universe. It not only provided support for the Big Bang theory, it also provided an understanding of the physics of the early universe that enabled predictions of the distribution of matter from the microwave background radiation forward and backward in time.

The discovery was made using a new statistical method, the false discovery rate (known as the FDR), to detect the oscillations. At false discovery rate 1/4, eight were flagged as possibly inconsistent with a smooth, featureless power spectrum. This and further analyses led the authors to conclude that the oscillations were statistically significant departures from a featureless matter-density power spectrum.

The method was developed through collaboration with two statisticians and published in *The Astronomical Journal*. Using this method, the authors were able to make their discovery and publish it in *Science* while other competing groups were still plowing through the plethora of data.

It is interesting to trace the history of this success, as it illustrates quite well how the "information hub" operates. Figure 4.3 illustrates the migration route of the statistical idea

When one tests many hypotheses on the same dataset, one must adjust the significance levels of the tests to avoid spurious rejection of true null hypotheses. This "simultaneous inference" problem has perhaps received the most attention in medical statistics – at least, all of the references cited as motivation appeared in the medical literature. Indeed, the main statistical contribution here was not to propose the sequential P-value procedure that was used in this example per se, which actually went back to Simes in the 80's (and maybe earlier), but rather to establish a convincing theoretical justification. This theoretical justification, the FDR control, led other researchers to propose a version for estimation.

The estimation proposal caught the attention of others because of its potential for threshold selection in wavelet shrinkage methods for statistical signal processing. Statisticians at CMU began work on FDR, both as a core statistics topic, and also in their collaboration with astrophysicists Miller and Nichol. Initially, they considered signal detection problems in huge pixel arrays. Later in their collaboration, the physicists recognized that this approach would apply to the acoustic oscillation signatures, which led

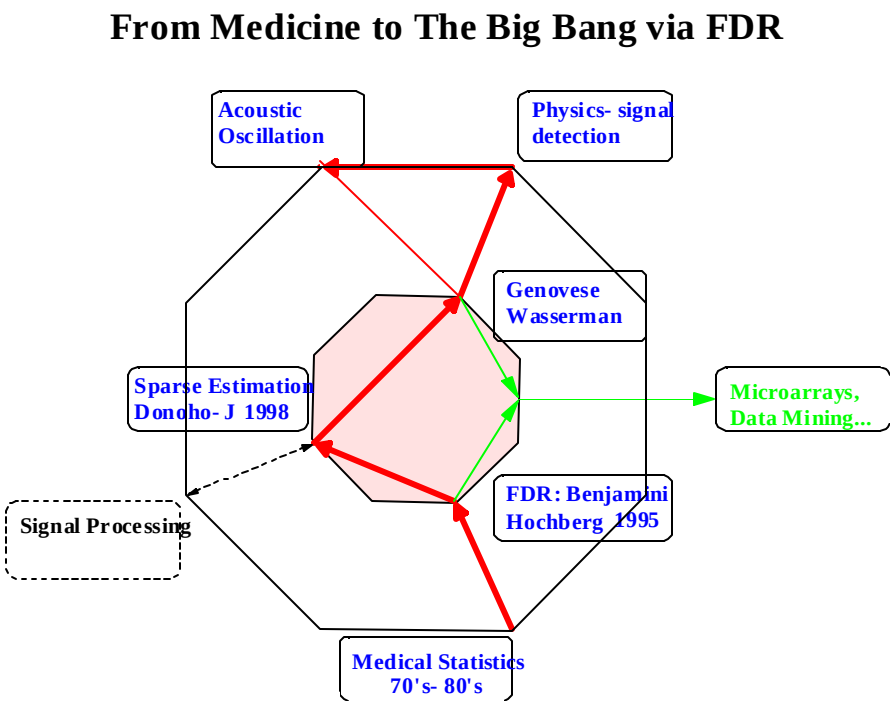

FIGURE 4.3. Migration of statistical ideas into subject matter areas

to the Science article.

Miller and Nichol report that when they give talks to the physics community on this work, there is great interest in the FDR approach. CMU physics professor Bob Nichol writes, in part "I personally would like to emphasize the symbiotic relationship that has grown between the Statisticians and astrophysicists here at CMU. It is now becoming clear that there are core common problems both sets of domain researchers find interesting e.g. application of FDR to astrophysical problems.

In fact, the astrophysicists appreciate the mathematical beauty of the statistics (and want to be involved), while the Statisticians clearly relish their role in helping to understand the Cosmos. In addition to these joint projects, this collaboration also is driving separate new research in the individual domains. In summary, this multiway collaboration has simulated both new joint research, as well as new separate research in the domain sciences. Therefore, it is a perfect marriage!

4.3 A set of research challenges

We next wish to suggest some of the important challenges that will be faced in future core area statistical research. To identify such challenges in statistics is inherently a slightly different enterprise than occurs in some other sciences. Whereas in mathematics, for example, much focus has been given to famous lists of problems whose challenge is enduring, in statistics the problems are always evolving corresponding to the development of new data structures and new computational tools. Unlike the laboratory sciences, statistics does not have big expensive problems with multiple labs competing–or cooperating–on major frontiers. It is perhaps more true in statistics than in other sciences that the most important advances will be unpredictable.

For this reason we need to maintain an underlying philosophy that is flexible enough to adapt to change. At the same time it is important that this future research should not degenerate into a disparate collection of techniques.

One can identify some general themes driving modern core area research. The challenges are based on the development of conceptual frameworks and appropriate asymptotic approximation theories for dealing with (possibly) large numbers of observations with many parameters, many scales, and complex dependencies.

The following subsections identify these issues in more detail.

4.3.1 Scales of data

It has become commonplace to remark on the explosion in data being gathered. It is trite but true that the growth in data has been exponential, in data analysts quadratic, and in statisticians linear. Huber's 1994 taxonomy of data sizes,

$$\text{Tiny } 10^2, \text{ Small } 10^4, \text{ Medium } 10^6, \text{ Large } 10^8, \text{ Huge } 10^{10}$$

already looks quaint (Wegman, 1995). For example, a single database for a single particle physics experiment using the "BaBaR" detector at the Stanford Linear Accelerator Center has 5×10^{15} bytes.

There will continue to be research issues at every scale – we haven't solved all problems for data sets under 100. However, a new part of the challenge to statistics is that the mix of issues, such as generalizability, scalability, and robustness, as well as the depth of scientific understanding of the data, will change with scale and context. Moreover, it is clear that our research and graduate training has yet to fully recognize the computational and other issues associated with the larger scales.

4.3.2 Data reduction and compression.

We need more "reduction principles": R. A. Fisher gave us many of the key ideas, such as sufficiency, ancillarity, conditional arguments, transformations, pivotal methods, and asymptotic optimality. Invariance came along later. However, there is clear need for new ideas to guide us in areas such as model selection, prediction, and classification.

One such idea is the use of "Compression" as a guiding paradigm for data analysis. The basic idea is that good structural understanding of data is related to our ability to compactly store it without losing our ability to "decompress" it and recover nearly the original information. For example, in the domain of signal and image data, wavelets are actually not optimal for representing and compressing curved edges in images. This suggests the need for new representational systems for better compression.

4.3.3 Machine learning and neural networks

Many ad hoc methods and computational strategies have been developed for "industrial strength" data. For the most part these methods are not informed by a broader understanding and integration into mainstream statistics. Thus future research should involve coherently integrating the many methods of analysis for large and complex data sets being developed by the machine learning community and elsewhere into the core knowledge of statistics.

This research will presumably be based on the building of models and structures that allow description of risk as well as its data-based assessment. It will include developing principled tools for guided adaptation in the model building exercise.

4.3.4 Multivariate analysis for large p, small n.

In many important statistical applications there are many more variables (p) than there are units being measured (n). Examples include analysis of curve data, spectra, images, and DNA micro-arrays. A recent workshop titled "High dimensional data: p>>n in mathematical statistics and in biomedical applications" in Leiden, Netherlands highlighted the current research importance of this subject across many areas of statistics.

The following more specific example can be offered to illustrate how innovations in other fields might prove useful in this problem, thereby reinforcing the idea that the core continually looks outward for ideas. Random Matrix Theory describes a collection of models and methods that have developed over the last forty years in mathematical physics, beginning with the study of energy levels in complex nuclei. In recent years these ideas have created much interest in probability and combinatorics.

The time now seems ripe to apply and develop these methods in high

dimensional problems in statistics and data analysis. For example, scientists in many fields work with large data matrices (many observations (n) and many variables (p)) and there is little current statistical theory to support and understand heuristic methods used for dimensionality reduction in principal components, canonical correlations etc.

Early results suggest that large n - large p theory can in some cases yield more useful and insightful approximations than the classical large n - fixed p asymptotics. For example, the Tracy-Widom distribution for "Gaussian orthogonal ensembles" provides a single distribution, which with appropriate centering and scaling provides really quite remarkably accurate descriptions of the distributions of extreme principal components and canonical correlations in null hypothesis situations.

4.3.5 Bayes and biased estimation

The decade of the nineties brought the computational techniques and power to make Bayesian methods fully implementable in a wide range of model types. A challenge for the coming decades is to fully develop and exploit the links between Bayesian methods and those of modern nonparametric and semiparametric statistics, including research on the possible combination of Bayesian and frequentist methodology.

One clear issue is that for models with huge data problems with large numbers of variables, the ideas of unbiasedness or "near" unbiasedness (as for the MLE) become less useful, as the idea of data summarization implicit in statistical methodology becomes lost in the complexity and variability of any unbiased method. This points to the need for a more extensive "biased estimation theory" and new theories for huge data problems with large numbers of variables.

Given their ever increasing use in all kinds of model-building exercises, it is also clear that there is a need for further analysis of "Monte Carlo" methods for inference.

4.3.6 Middle ground between proof and computational experiment.

A final challenge for theoretical work in the coming decades is to develop an agreed-upon middle ground between the pace of proof (too slow), and the swamp of unfettered computational experimentation (too arbitrary and unconvincing). There are many problems in which rigorous mathematical verifications might be left behind in the development of methodology both because they are too hard and because they seem of secondary importance. For example, despite many years of work, there are important families of statistical models, such as mixture models, in which identifiability questions are largely ignored because of the difficult analysis that is involved and the

ever-widening variety of model structures that must be investigated.

4.4 Opportunities and needs for the core

If there is exponential growth in data collected and in the need for data analysis, why is core research relevant? It is because unifying ideas can tame this growth, and the core area of statistics is the one place where these ideas can happen and be communicated throughout science. That is, promoting core area statistics is actually an important infrastructure goal for science from the point of view of efficient organization and communication of advances in data analysis.

A healthy core of statistics (through a lively connection with applications) is the best hope for efficient assimilation, development and portability between domains of the explosion of data analytic methods that is occurring. As such, it is a key infrastructure for science generally.

4.4.1 Adapting to data analysis outside the core

The growth in data needs provides a distinct challenge for statisticians to provide, in adequate time, intellectual structure for the many data analytic methods being developed in other arenas. As one leading statistician said, "If we don't want to be taken over by Machine Learners or Computer Scientists, people who work exclusively in some interesting area of applications, and have a natural advantage on their own turf, we have to keep thinking of good statistical ideas, as well as making them friendly to the users."

4.4.2 Fragmentation of core research

It is our perception that statistical outreach activity is high and it is increasing for all sorts of good reasons. Unifying ideas can tame this growth, and the core of statistics is the one place where these ideas can happen and be communicated throughout science. But there has been what we think is an unintended consequence of this growth – a relative neglect of basic research, and an attendant danger of our field fragmenting.

We emphasize again the importance of core research: the FDR example illustrates that methodologic/theoretical insight into ad hoc methods magnifies their potential for application.

One might mention some data items to support this: In previous years, according to an "export scores" analysis by Stephen Stigler, the Annals of Statistics was the most influential statistics journal. However, reflecting recent trends, submissions to this journal are down by about 25%, and perhaps not co-incidentally, the fraction of US authors has dropped from 70% twenty years ago to 35% now.

This manpower problem is destined to grow worse, as it is clear that Ph.D. students in statistics are finding, through the job market, that outreach skills are highly valued.

4.4.3 Growth in the professional needs.

The core research of statistics is multidisciplinary in its tools: it borrows from (at least) information theory, computer science, and physics as well as from probability and traditional math areas.

As statisticians have become more and more data-focused (in the sense of solving real problems of modern size and scope), the math skills needed in core areas have gone up. To name a few areas, statistician might need to know complex analysis (saddlepoints), algebra (contingency tables), Markov chains (MCMC), or functional analysis (complex model building). At the same time, there is the need to be enough of a computer scientist to develop the algorithms and computer software required for the data analysis.

This need for ever increasing technical skills provides yet a second set of challenges to keeping the core vital as a place for integration of statistical ideas.

4.4.4 Research funding

It seems clear that funding for core research has not kept pace with the growth of the subject. Investigators, rather than beating their heads against difficult funding walls, turn their efforts towards better funded outreach activities or consulting.

But the whole history of statistics shows that inreach activity in statistics is critical to its own health, and highly leveraged in its payoff across the sciences. Many statisticians are concerned about the current levels of support.

It is therefore suggested that we raise the profile of support for core research to counteract the hollowing out of external sources. The most basic needs remain as they always have: to encourage talent, giving senior people time and space to think, and encouraging junior people to buy into this line of research.

4.4.5 A Possible Program

One might ask, what sort of special program might a funding agency offer that would stimulate the integrative activity of the core? In that spirit, we provide a specific suggestion for a program along with the possible name : "Method Exploration and Migration". It represents one possible way to enhance the ability of researchers to provide integration of new ideas arising outside the intellectual core.

The theme of the program would be "developing new data analytic ideas for broad scientific use". It would be based on the observation that most data analytic innovation necessarily occurs in a particular scientific context, but is potentially of far wider applicability. It might support research to understand common features of (new) data analytic techniques in fields A (and maybe B, C) with a view to understanding their properties and promoting their use in science generally. Such a program need not be restricted to statisticians.

There would be some subtle differences in this program compared with the usual interdisciplinary research initiatives. Interdisciplinary research typically brings together researchers from fields A, B and C to collaborate on grand challenge D. This research might also require statisticians (or other methodologists) in a collaboration with researchers in A (and maybe B and C) to understand the context and current uses of the techniques.

The key difference in the new program would be that the primary goal would not be to advance the use of techniques existing in field A per se (though this may be a positive side effect). The aim rather is through further research to understand and explain in general terms why the methods may be of wide utility. The point is that this work might not take place otherwise, as the pace of work in hot field A is typically such that its scientists and even its methodologists have neither time nor support to promote the advance of data analysis generally.

How does this differ from standard "investigator-initiated grants"?–here support would be needed for the statistician/methodologists to develop familiarity, contacts, and collaborations with field B and then to conduct the research. This is not part of the usual grants. But field B shouldn't be asked to fund the specific project, since it isn't the primary beneficiary. It is an inreach or "methodologic infrastructure" activity.

Of course, it is possible to do such research now as historical examples in the development of core statistics attest. But giving this activity higher profile and support would offer high leverage benefits to the data analytic arsenal of science generally.

5

Statistics in Science and Industry

A distinguishing feature of statistics as a discipline is its interaction with the entire spectrum of natural and social sciences and with technology. This chapter is concerned with the elucidation of the role of statistics in gathering knowledge across a wide spectrum of possibilities.

5.1 Biological Sciences

Building on the foundations of agricultural and genetic statistics developed in the first half of the 20th century, biostatistics, statistical epidemiology, and randomized clinical trials have been cornerstones of the systematic attack on human disease that have dramatically increased life expectancy in advanced societies during the past half century.

Recent progress in molecular biology and genetics has opened entirely new areas of investigation, where for the foreseeable future there will be rapid advances in understanding fundamental life processes at the molecular level. The long term goals of this research are the application of the knowledge of molecular processes to entire organisms and populations. These goals include improved tailoring of medical treatments to the individual (e.g., by devising treatment suited to the individual's genetic makeup), alleviation of problems of malnutrition and starvation by improving agriculturally important plant species and domestic animals, improved public health, and better defense against bioterrorism.

In addition to new solutions for problems that arise out of the "new" biology discussed below, success in statistical research will also depend on better understanding and further development of the statistical methods for clinical trials, laboratory and field experiments, and observational studies that have been developed during the past half century.

At the risk of oversimplifying the many new developments in biological research, it is useful to consider four areas where statistical and computational methods have played and will continue to play an important role: (A) computational genomics, including in particular biomolecular sequence analysis and functional genomics, (B) genetic epidemiology and gene mapping, (C) evolution, population genetics, and ecology; and (D) computational neuroscience.

(A) **Biomolecular sequence analysis and functional genomics** re-

fer to methods based on analysis of DNA sequences (the building blocks of genes) and amino acid sequences (the building blocks of proteins), and global profiles of RNA and proteins in various cellular states, to discover the structure and evolution of genes and proteins, and their functions in normal and abnormal processes. Examples include

1. data base searches based on protein sequence alignment to infer functions of a newly discovered protein by comparing it with possibly related proteins that have already been studied,

2. the identification of control regions imbedded in the genome that govern the amount of protein produced and the conditions under which it is produced,

3. alignment of homologous genomic regions of different plant or animal species as a first step in inferring their phylogenetic relationships, and

4. comparative analysis of the levels of gene expression in normal and diseased cells to provide objective differential diagnostics for diseases that present similar clinical symptoms and, more ambitiously, to provide avenues for successful treatment based on understanding the role of the over and under expressed genes in the pathology of the disease.

Promising new directions in this area include the use of computational and functional genomics approaches in areas such as molecular medicine and cellular and developmental biology.

Molecular medicine seeks to use genetic data to identify subjects at risk for drug toxicity, to develop refined classification of disease subtypes based on genotype, RNA and protein profiles, and to develop individualized therapeutic intervention based on predictive models that use molecular level assays. Justification of research in this direction will ultimately depend on traditional clinically oriented biostatistical areas such as clinical trails and cohort studies. This is an area of unlimited opportunities for the discipline of biostatistics.

Although statistics has not yet been firmly established in **cellular and developmental biology**, it appears that new statistical and computational approaches will be essential for future advances as more and more high throughput experimental approaches are designed, e.g., recently implemented assays in 96 or 384 well format to obtain real time measurements of the activities of thousands of gene promoters in parallel.

An astounding amount of imagery based on time-lapsed microscropy, in situ hybridization and antibody staining will provide a dynamic view of key molecular events at every stage of an organism's development. One particularly exciting direction is the development of approaches that are capable of integrating information from primary literature (PubMed, online articles) and knowledge bases (e.g. Locus Link, OMIM, Flybase, Gene

Ontology), with the analysis of high-throughput functional genomics and cellular imaging data.

(B) The goal of **Genetic epidemiology** is to understand the relative importance of environment and genetics in human disease. **Gene mapping** involves the use of maps of molecular markers throughout the genome of a particular plant or animal to locate the genes that contribute to phenotypes of interest. It is frequently the first step toward better understanding and treatment of those diseases in plants and animals where inheritance plays an important role. One also wants to map genes that lead to desirable traits in agriculturally important plants and domestic animals or genes in model organisms like the laboratory mouse that may provide clues to the genetics of similar human phenotypes.

In experimental organisms genetic mapping includes the design of breeding experiments to maximize information. Gene mapping in humans, where one cannot perform breeding experiments, is much more complex, with some approaches exploiting relationships within families, while others involve the more difficult to infer and more complex relationships of individuals within populations.

(C) **Evolution, population genetics and ecology** study the changes that occur at the population level in plants and animals in response to random mutational changes in the population's gene pool and changes in their environment. Although originally oriented toward the study of evolutionary relationships (for example, the evidence supporting the hypothesis of a common African origin of modern humans), the ideas of population genetics are increasingly used to understand the evolution of bacteria and viruses (in order to provide appropriate vaccines and drugs) and the evolution of proteins in different species of plants and animals (in order to understand protein structure and function by identifying parts of related proteins in different species that have been conserved by evolution).

(D) Using modern methods of neuroimaging (PET, fMRI), **computational neuroscience** attempts to understand the functioning of nervous systems at the level of small numbers of interacting neurons and at the level of the entire brain: which parts of the brain are activated under which conditions? How do the brains of normal and psychotic individuals differ in their structure and/or function? How can we use this knowledge for diagnosis and treatment?

Computational neuroscience encompasses basic molecular biology from the study of ion-channel behavior, modeling of neuronal firing in simple networks, and responses of olfactory and visual receptors, to macroscopic measurements using in vivo brain imaging and cryo-sectioning techniques, to abstract approaches to computational vision. Statistics plays a vital role at each level of analysis.

Statistical and Computational Methods

As a consequence of this enormous diversity of scientific problems, a

expansive set of statistical, probabilistic, and computational methods has proved to be very useful. Some methods have proved themselves in a number of areas, while others have more specialized applications.

Stochastic processes, from finite Markov chains to point processes and Gaussian random fields, are useful across the entire spectrum of problems. Statistical techniques of classification, clustering, and principal components are widely used in (A) and (D). Likelihood and/or Bayesian analysis of stochastic processes is important in (A), (B) and (C). Because of the large amount of data produced, e.g., expression levels on a microarray for tens of thousands of genes in a sample of individuals, or data from up to a thousand markers (in the future perhaps one hundred thousand) distributed across the genome of thousands of individuals, challenging issues of multiple comparisons arise in (A), (B) and (D).

Hidden Markov models and Markov chain Monte Carlo provide important computational algorithms for calculating and maximizing likelihood functions in (A), (B), and (C). Some of these statistical methods are classical (e.g., principal components, likelihood analysis), but even they may require adaptation (principal curves, likelihood analysis of stochastic processes) to deal with the large amounts of data produced by modern biological experiments. Other methods (hidden Markov models, Markov chain Monte Carlo) have developed relatively recently in parallel with the modern computing technology necessary to implement them.

In addition there are some methods that are of paramount importance to the development of a single area. An example is the use of trees (phylogenetic and coalescent trees) to describe evolutionary relationships among individuals within a population and among different populations. (Trees also play a technical role in cluster analysis.) Experimental design and variance components provide important tools for genetic mapping.

Many techniques have been developed in close relation to the field of application, and it is expected that important contributions in the future will come from statisticians who are well versed in specific applications. However, even these techniques have typically built on a theoretical structure that was developed earlier in response to some other perceived need, often in a field far removed from modern biology.

The common methodological features of those methods that find application in several different areas provide motivation to achieve better theoretical understanding, even when that understanding is not tied to a specific application. It is also worth noting that in view of the vast explosion of knowledge, much of it cutting across traditional disciplinary lines, training the next generation of scientists will require some (not yet apparent) consensus about what concepts will be important and the appropriate balance between general methodology and specific subject matter knowledge.

A common feature of all the efforts described above is the amount, complexity and variability of data, with the result that computation (frequently including graphics) is an important aspect of the implementation of every

idea. In view of the diverse mathematical and computational backgrounds of scientists engaged in biological research, it is important that computational algorithms be made as "user friendly" as possible. This may require support for specialists to provide the "front end" and documentation necessary so that laboratory scientists can use tools developed by statisticians easily and correctly.

Illustrative Examples

Example 1. A subject showing the importance of a broad mathematical viewpoint in solving concrete problems of biological importance is found in the assessment of statistical significance in gapped sequence alignments (referred to under (A)1 above).

The modern history of this subject began about 1990, when a team of researchers for the purpose of analyzing *single* DNA or protein sequences, recognized the relevance of results obtained by Iglehart in 1970, in his investigations of queueing theory. At the same time others conjectured that a similar result would hold for pairwise sequence alignments, a much more difficult result that was proved by another team of researchers in 1994, but only for the artificially simplified problem of alignments without gaps.

Based on conjectures of Karlin and Altschul and of Waterman and Vingron (1994) that an approximation of the same parametric form would be valid for the more important practical case of gapped alignments, Monte Carlo methods were developed to estimate the parameters of the conjectured approximation. These Monte Carlo estimates have been encoded into the widely used BLAST software, but their application is limited to a small number of previously studied cases by the slowness of the required computations.

Using methods motivated by applications to quality control, an approximation was obtained for gapped alignments that is much more easily evaluated, albeit less accurate. Current research continues in an attempt to find an approximation that successfully combines generality, speed of evaluation, and accuracy.

Example 2. An area that has stimulated rapid development of new computational and statistical tools is the analysis of cDNA microarrays, which are used for measuring gene expression in a wide variety of biological studies. A typical problem is to assess differential expression between a control and a treatment group for a large number (thousands) of genes from a relatively small sample of individuals. Descriptive statistics, often in the form of clustering algorithms, and inferential statistics to deal with special problems arising from the simultaneous comparison of thousands of genes both play important roles. For example, the collaboration of statisticians and researchers in oncology and biochemistry produced the software "Significance analysis of microarrays" (SAM) (Stanford University). This development was motivated by an experiment to measure the genetic response of human cells to ionizing radiation. The method is very simple, and

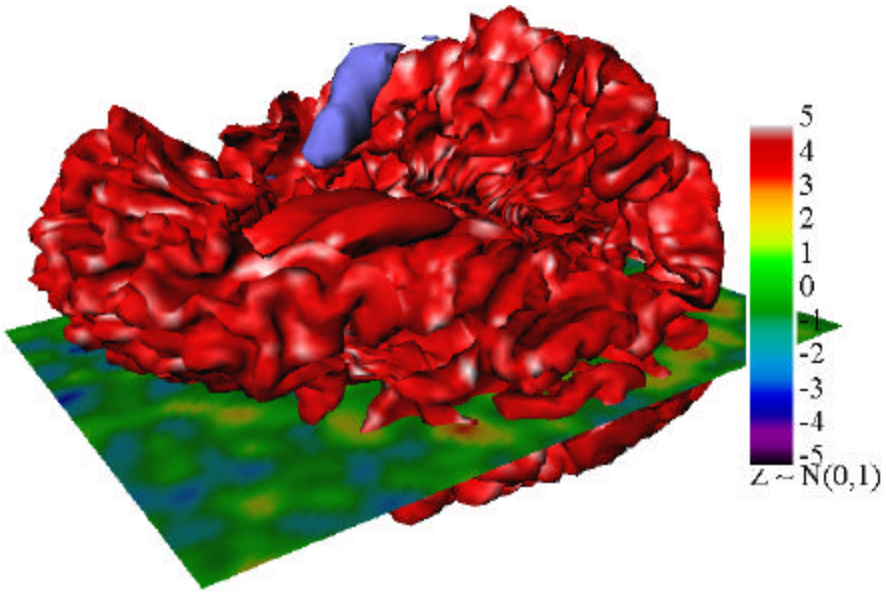

FIGURE 5.1. An illustration of brain imaging from Example 3

was implemented as an Excel spreadsheet add-in "SAM." This package has been downloaded over 3400 times since its release in April 2001.

Example 3. An example illustrating the importance of both higher mathematics and of computational methods to promote visual understanding of complex data is provided by the research of K. Worsley, who for most of the last decade has focused on brain imaging data obtained either from positron emission tomography (PET) or functional magnetic resonance imaging (fMRI) (and similar astrophysical data) (cf. Worsley, Evans, Marrett and Neelin (1992)) or Worsley, *et al.* (2002). Worsley has used mathematical ideas of differential and integral geometry developed by pure mathematicians beginning with C. F. Gauss in the 1800s, in order to assess the statistical significance of regions of apparent neural activity in response to particular external stimuli. An example of the accompanying graphics derived from his software is given in Figure 5.1.

In summary, the large amounts of data produced by modern biological experiments and the variability in human response to medical intervention produce an increasing demand for statisticians who can communicate with biologists and devise new methods to guide experimental design and biological data analysis.

5.2 Engineering and Industry

Historical Perspective and Recent Developments

Statistical concepts and methods have played a key role in industrial development over the last century. Applications in engineering and industry, in turn, have been major catalysts for research in statistical theory and methodology. The richness and variety of these problems have greatly influenced the development of statistics as a discipline.

The origins of industrial statistics can be traced to the pioneering work of Walter Shewhart on statistical process control (SPC) in the 1920s. Today, SPC philosophy and methods have become a critical technology for quality improvement in manufacturing industries and are increasingly being used in business and the service and health industries.

The early work on design of experiments (DOE) by R. A. Fisher, F. Yates, and their collaborators at Rothamsted Experimental Station was stimulated by the needs of the agricultural industry. Product testing, analysis, and optimization in the chemical and textile industries led to further developments in factorial designs and new methods such as response surface methodology and evolutionary operation by G. Box and others.

The emphasis on quality improvement and G. Taguchi's ideas on robust design for variation reduction have led to extensive research in and application of designed experiments for product and process design, quality and process improvement, and software testing. The needs of the defense, electronics, aircraft, and space industries have also stimulated the development of new areas such as sequential analysis, reliability, spectrum estimation and fast Fourier transform algorithms.

The years during World War II saw rapid growth of the use of statistical methods in quality control. After a period of stagnation, the renewed focus on quality and productivity improvement during the last three decades has rekindled interest in and appreciation for statistics in industry. Statistical concepts for understanding and managing variation and basic statistical techniques of the DOE and SPC form the backbone of popular quality management paradigms such as Total Quality Management (TQM), Six Sigma, and Deming's famous 14 points. Major companies have invested heavily in re-training their workforce in quality management principles and basic statistical methods.

Much of the early work was driven by the needs of the agricultural, manufacturing, and defense industries. In recent years, the scope has expanded substantially into business and finance, software engineering, and service and health industries. Applications in these areas include credit scoring, customer profiling, design of intelligent highways and vehicles, e-commerce, fraud detection, network monitoring, and software quality and reliability.

While the benefits are hard to quantify, it should be clear from even this abbreviated historical summary that statistics and statistical thinking have

had a profound positive impact on engineering and industry in the United States.

High Impact Research Areas

Global competition and increasing customer expectations are transforming the environment in which companies operate. These changes have important implications for research directions in statistics. Following are brief descriptions of four general examples.

A. Massive Data Sets with Complex Structure: This topic cuts across all parts of business and industry (as well as other areas discussed in this report). Business and manufacturing processes are becoming increasingly complex. Consequently, engineers and managers are in greater need of relevant data to guide decision-making than ever before.

At the same time, advances in sensing and data capture technologies have made it possible to collect extensive amounts of data. These data often have complex structure in the form of time series, spatial processes, texts, images, very high dimensions with hierarchical structure, and so on. Collection, modeling, and analysis of these data present a wide range of difficult research challenges.

For example, monitoring, diagnosis, and improvement of advanced manufacturing processes require new methods for data compression and feature extraction, development of intelligent diagnostics, and real-time process control. These problems also involve issues of a general nature such as selection biases, computing, scalability of algorithms, and visualization. Statisticians have important roles to play in designing effective data warehousing solutions, ensuring data quality, developing informative data collection and data reduction (compression) schemes in this new environment. Many of these issues have, until recently been dominated by computer scientists and engineers.

To be effective, however, the methods must be developed in the context of specific applications, and the empirical information must be integrated with engineering and subject-matter knowledge for decision making. For example, a research project on yield improvement in semiconductor manufacturing led to new methods for analyzing and visualizing spatial data, including methods for monitoring spatial processes, characterizing spatial patterns, and development of fault diagnostics. Engineering research on stamping processes resulted in new methods for monitoring functional data combining wavelet techniques with engineering knowledge for data compression and feature extraction.

Other areas of application such as credit scoring, fraud detection in telecommunications, and warranty analyses are also generating many research problems. Warranty costs in the automobile industry now run into billions of dollars annually (not counting safety and lives lost). There is a need for methods that quickly detect warranty problems (small signals) from very large and extremely noisy data sets.

Much of the past work has also focused on individual processes without taking a holistic approach to modeling and optimization. One of the grand challenges is the need for enterprise-level modeling and to "instantaneously transform information from a vast array of diverse sources into useful knowledge and effective decisions."

B. Large-Scale Computational Models – Experimentation, Analysis and Validation: Computational models and simulation are being used more and more frequently in many areas of application. In manufacturing industries, competitive market forces and the concomitant pressure to reduce product development cycle times have led to less physical testing and greater use of computer-aided design and engineering (CAD/CAE) methods. Finite-element analysis and other techniques are used extensively in the automobile industry for product design and optimization.

There are similar trends in semiconductor manufacturing, aircraft, defense, and other industries. The computational models are very high dimensional, involving hundreds and even thousands of parameters and design variables. A single function evaluation can take several days on high-end computing platforms.

Experimentation, analysis, visualization, and validation using large-scale computational models raise a variety of statistical challenges. These include: a) development of experimental designs for approximating and exploring response surfaces in very high dimensions, b) incorporating randomness and uncertainty in the design parameters and material characteristics into the computational model; c) modeling, screening, prediction, and optimization.

There has been some research on design and analysis of computer experiments in the literature, including the development of new classes of designs and the use of Gaussian random fields and spatial interpolation techniques for inference (National Research Council (1996)). But research in this area has not kept pace with the needs of industry. Validation of large-scale computational models has received relatively little attention in the statistical literature. Sequential methods, DOE, and Bayesian analysis, among others, have important roles to play here. There are also opportunities for collaboration with researchers in numerical analysis and optimization.

C. Reliability and Safety: The design, development and fabrication of highly reliable products that also meet safety and environmental goals represent another area of major challenge faced by industry. The traditional focus in reliability has been on the collection and analysis of "time-to-failure" data. This poses difficulties in high-reliability applications with few failures and high degrees of censoring.

Fortunately, advances in sensing technologies are making it possible to collect extensive amounts of data on degradation and performance-related measures associated with systems and components. While these data are a rich source of reliability information, there is a paucity of models and methods for analyzing degradation data and for combining them with physics-

of-failure mechanisms for efficient reliability estimation, prediction, and maintenance. Degradation analysis and device-level failure prediction are integral parts of predictive maintenance for expensive and high-reliability systems.

New, modern materials being developed, such as various types of composites or nanostructured materials, require research on appropriate models and methodology for prediction of failure and other properties. Modern aircraft and other structures will increasingly use these materials for critical parts whose failure could be catastrophic, bringing user safety to the forefront. Statisticians will need to work closely with materials scientists and engineers to be successful in this arena.

There are also vast amounts of field-performance data available from warranty and maintenance databases. Mining these data for signals and process problems and using them for process improvement should be a major area of focus. There is also a need to incorporate the environment in which a system operates into reliability models and analysis of field-performance data. These environments are generally dynamic and/or heterogeneous, and development of realistic models for reliability assessment and prediction under such conditions will be needed.

D. Software engineering: This is still a relatively new field when compared with traditional branches of engineering. Its importance to the nation is underscored by the increasing reliance of the U.S. economy and national defense on high quality, mission critical software (National Research Council (1996)).

Statistics has a significant role to play in software engineering because data are central to managing the software development process, and statistical methods have proven to be valuable in dealing with several aspects of it. To mention a few examples, statistical considerations are essential for the construction and utilization of effective software metrics, and experimental design ideas are the backbone of technology for reducing the number of cases needed to test software efficiently (but not exhaustively). Further, statistical quality control provides the basis for quantitative analysis of various parts of the software process and for continuous process improvement.

Indeed, the entire movement towards formal processes for software development, as in the Software Engineering Institute's Capability Maturity Model, can be traced in part to the pioneering work of W. A. Shewhart and W. E. Deming on statistical quality control and related topics. In spite of the progress that has been made, considerable additional research will be essential to deal with the software challenge (or more dramatically the "software crisis").

5.3 Geophysical and Environmental Sciences

Background

The term 'geophysical and environmental sciences' covers many specific fields of study, particularly if environmental sciences is taken to include the study of ecological phenomena and processes. This broad area of statistical activity does not have an easily summarized history, nor a simple pattern of development. Indeed, the history of statistical work in the geophysical and environmental sciences is intertwined with fields as diverse as agriculture, basic biology, civil engineering, atmospheric chemistry, and ecology, among others.

Rather than give a broad and necessarily incomplete survey of areas in which statistics has had, and continues to have, an impact, this presentation focuses on topics that illustrate several aspects of the interplay between statistics and other scientific disciplines. In particular, examples have been chosen to illustrate the tandem use of deterministic process models and stochastic models, the use of models for correlated data in the detection of change in environmental processes, and the role of statistical thinking in scientific conceptualization.

Deterministic Process Models and Stochastic Models

A substantial amount of emphasis is now being placed on the tandem use of deterministic process models and statistical models. Process models have typically taken fundamental scientific concepts such as mass balance in chemical constituents as a foundation, and built up more elegant mathematical structures by overlaying equations that represent physical and chemical interactions, often in the form of sets of differential equations. Statistical models, on the other hand, typically rely on the description of observed data patterns as a fundamental motivation for model development. Increasingly, there is recognition that ones understanding of many geophysical and environmental processes can be advanced by combining ideas from these two modeling approaches.

One method that has been used to combine process and statistical models is to use the output of deterministic models as input information to a stochastic formulation. An example, is the analysis of bivariate time series representing northern and southern hemispheric temperature averages. Along with traditional use of linear trend terms and covariate information such as effects of the El Niño-Southern Oscillation (ENSO) phenomenon, the outputs from 24 deterministic climate models were considered for their ability to describe hemispheric mean temperatures over the period 1900 to 1996.

As an example of the results, Fig. 5.2 shows the raw data for both hemispheres with both a fitted straight line and the estimated trend curve arising from a combination of climate forcing factors (i.e., climate model output) and El Niño-Southern Oscillation effects. It is visually obvious, and the

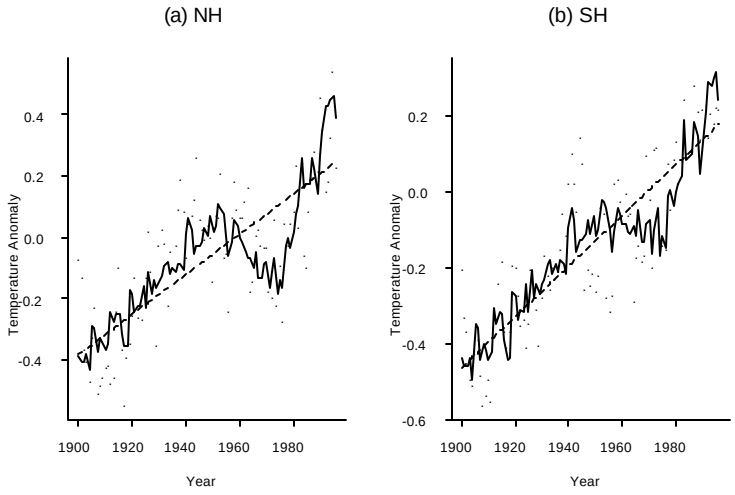

FIGURE 5.2. Observed temperature anomalies for northern (NH) and southern (SH) hemispheres, with fitted straight lines (dashed lines) and estimated trends (solid curves) due to combination of climate forcing factors and ENSO. From Smith *et al.* (2001).

detailed statistical analysis confirms, that the latter trend curve fits the data much more closely than a simple straight line regression. This example demonstrates the use of a statistical analysis to provide evidence about which factors are important for inclusion in process models, as well as providing a description of observed data.

Another method for incorporating statistical and deterministic modeling approaches is illustrated by recent work on ocean surface wind speeds authored by statisticians associated with the Geophysical Statistics Project at the National Center for Atmospheric Research (NCAR) and an oceanographer. Data were obtained from two sources. One is satellite data that arise from the NASA scatterometer (NSCAT) instrument. These data are high-resolution but of sparse spatial coverage. The second source of data is so-called *analyses* produced by a global-scale numerical weather prediction model of the National Center for Environmental Prediction (NCEP). These data are complete in the sense that each six-hour observation covers the whole region, but is of much lower spatial resolution than the NSCAT measurements.

Statistical analysis of these data requires techniques beyond standard spatial and spatio-temporal statistics. In addition to both temporal and spatial components, the analysis must accommodate several sources of data that are dissimilar in resolution and scales of coverage, with a goal of pro-

viding a faithful representation of wind speed over the entire region for each six-hour period.

Correlated Data and Environmental Trends

Many environmental problems involve the detection and estimation of changes over time. For example, an environmental monitoring agency such as the EPA uses trend estimates to assess the success of pollution control programs and to identify areas where more stringent controls are needed. In climate modeling, a major preoccupation is to determine whether there is an overall trend in the data, not only for widely studied variables such as the global mean temperature but also for numerous other variables where the conclusions are less clear cut.

For statisticians, the estimation of trend components with correlated errors is a problem with a long history, and much of this work involved substantial interaction between statisticians and geophysical and environmental scientists. For example, Sir Gilbert Walker, known to statisticians through his many contributions to time series analysis and in particular the Yule-Walker equations, was also a distinguished meteorologist who worked extensively on the El Niño-Southern Oscillation (ENSO) phenomenon, and these contributions were largely the result of the same research.

A long collaboration between statisticians and geophysicists has resulted in a series of papers on the detection of change in stratospheric ozone in which a large number of models with correlated errors are considered. This research, consisting largely of papers with a statistician as lead author but appearing in journals outside of the mainstream statistical outlets, is an excellent illustration of the outreach of statistics to other scientific fields.

Numerous authors working on problems from the atmospheric sciences have also considered models with correlated errors and, in particular, have examined how the conclusions about climate change vary with different assumptions about the error process. Such assumptions include time series models with long-range dependence, and models using spectra derived from climate model dynamics. Other workers have presented an alternative approach using wavelet representations of long-range dependent processes, and continuing work in this area illustrates the feedback that consideration of the important scientific problem of climate change has on the development of new statistical representations for environmental processes. Recent work, authored by statisticians but published in the meteorology literature, has taken statistical models for long-range dependent processes developed largely in the analysis of problems from economics and applied them to wind speed and direction. This illustrates the role of statistics in the transfer of methodology from one discipline (in this case, economics) to another (meteorology) that may have otherwise remained unaware of its applicability to their problems.

Statistical Modeling and Scientific Conceptualization

It is common for changes in environmental data records to be conceptu-

alized within the statistical framework of *signal* plus *noise*. Indeed, this is the case for many of the models discussed above, in which various forms are given to the signal (or systematic) and noise (or error) components of the models to better represent the processes under study. Consideration of a signal plus noise structure in the analysis of water chemistry variables has led many limnologists to conclude that the data records observed contain a small signal component embedded in a large noise component.

But this concept can lead to difficulties, as illustrated by considering records of Secchi depth (a measure of water transparency) for Lake Washington in the United States. Figure 5.3 presents observed Secchi depths for a relatively short time span in 1980 to 1981. The panel on the left displays a sequence of values that clearly indicates an increasing trend. But, when these values are embedded in a slightly longer sequence of observations in the right panel, we can see that this change is not meaningful in terms of determining whether a substantial change has occurred in the basic character of the lake.

Nevertheless, it is difficult to accept that the values from mid-1980 to early 1981 are a realization of noise component alone. In this example, the environmental processes of interest may be better conceptualized as consisting of a number of *layers* of processes, each of which may contain its own noise component.

A model for this situation has been formulated as an extension of a Bayesian dynamic model. The model consists of three conceptual processes, an *observational* process, a *current condition* process, and a *lake function* process, which is of greatest interest in monitoring water quality. The model can be shown to clearly detect changes in the lake function process for Lake Washington at three times between 1960 and 1990; those times correspond to three known events that have impacted the basic character of Lake Washington.

Here, statistical methods have helped in the conceptualization of an environmental situation for which the development of process models is prohibitively complex. That is, the three processes used in the dynamic model do not correspond to physical or chemical mechanisms, but rather to scientific conceptualization of environmental processes, in a manner similar to that of the fundamental limnological concept of lake *trophic status*.

Another example in which statistics has aided in the development of scientific thinking is the analysis of cycles in the populations of Canadian lynx and snowshoe hare, and a series of papers dealing with this have appeared in the *Proceedings of the National Academy of Science* and *Science*. Here, collaboration between statisticians and ecological scientists has resulted in a strengthening of scientific theory. A number of concepts have been developed through this work, including the relation between autoregression order of a statistical model and the complexity of feedback systems between species (i.e., lynx and hare) and the idea that population cycles may exhibit spatial synchrony.

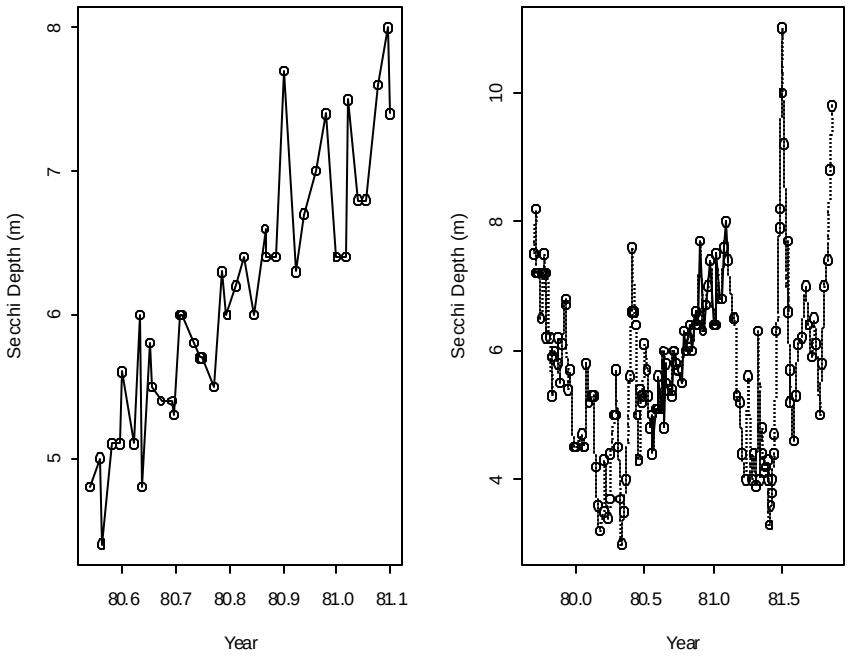

FIGURE 5.3. Observed Secchi depth values from Lake Washington. The graph of the left panel is embedded as the solid line on the right panel.

In particular, researchers analyzed 21 time series of lynx populations spanning the period 1821 to the 1990s. They employed nonlinear autoregressive processes of order 2, and combined series through random coefficients and empirical Bayes estimation. Having developed good statistical descriptions of the observed data, they then derived equivalent forms of pure mathematical models in theoretical population ecology.

Depending on one's perspective, the mathematical models of population dynamics give meaning to the statistical models used, or the statistical models have helped guide the development of theoretical descriptions of populations. This work resulted in a demonstration of the inter-relation of density dependence and phase dependence in population cycles.

Other directions and future possibilities

The collection and processing of large amounts of data is a feature in

many of the major components of the geophysical and environmental sciences such as meteorology, oceanography, seismology, the detection and attribution of climate change, and the dispersion of pollutants through the atmosphere.

Statisticians have been actively involved in all of these fields, but as statistical methodology has advanced to include, for example, complex models for spatial-temporal data and the associated methods of computation, the potential for direct interactions between statisticians and geophysical and environmental scientists has increased enormously. Traditional methods of multivariate analysis and spatial statistics rely heavily on matrix computations which are infeasible in very large dimensions; this has provoked research on methods which perform well in such situations and that are computationally efficient with large data sets.

Much statistical effort has recently been devoted to the development of models and methods appropriate for the analysis of large-scale spatial and temporal data sets; the model described above for ocean windspeed data is just one instance of new statistical methodology developed in response to these problems. Modeling approaches that are being developed for use in such situations include hierarchical forms of geostatistical mdoels and general mixture models based on lattice structures.

Another major area of research is on nonstationary spatial models including methods to represent a nonstationary process using a deformation of space that results in stationarity, models defined by kernel smoothing of an underlying random field, and models defined by multiresolution bases.

The design of monitoring networks is another area with a large literature. Design problems, such as the locations at which samples should be taken when it is possible to take only a small number of measurements in a large complex system, are also relevant in the context of data assimilation.

One expects to see a huge growth in the analysis of data from numerical environmental models, and direct interaction with applied mathematicians and subject-matter scientists in the development of such models. Models of the kind we are thinking about arise in climatology and numerical weather forecasting; the modeling of atmospheric pollutants; and modeling flow in a porous medium. This topic is the subject of forthcoming program at the Statistical and Applied Mathematical Sciences Institute (SAMSI) in 2003 and is likely to be a very large subject for future research.

Statisticians have also been active in addressing questions of ecological concern, although the fundamental statistical questions in ecological analyses are less evident than they are for modeling environmental processes in space and time. Thus, statistical contributions tend to be more scattered, and there are few long-term collaborative teams of ecologists and statisticians.

Historically, statisticians have made contributions to sampling problems in ecological field studies, the assessment of population estimation, and analysis of community composition. Funding sources for ecological work,

too, have been less available for statistical development in the ecological sciences than they have been for the analysis of atmospheric processes and of pollution fields.

With increased emphasis on the assessment of biodiversity as a far-reaching issue with both scientific and social implications this area is ripe for both application and stimulation of statistical research. The emergence of what is often called landscape ecology brings with it a broadening of spatial extent for the consideration of ecological questions and, here too, there is both a need and opportunity for increased statistical activity.

Funding needs and opportunities

DMS has supported a number of initiatives on statistics in the geophysical and environmental sciences. Examples include the three-year (1998–2000) joint initiative with the EPA; DMS's support since 1995 of the Geophysical Statistics Project (GSP) at NCAR, and the current GeoMath initiative; and the SAMSI program entitled "Large-scale computer models of environmental systems."

These initiatives are welcome and have greatly increased the potential for individuals trained as academic statisticians to work collaboratively with scientists in the geophysical and environmental sciences. Developing effective collaborative research programs requires a considerable investment of time and energy on the part of both statisticians and applied scientists.

There is the expectation that faculty who develop such collaborative relationships will be able to secure external funding to continue to support those efforts. Young faculty, in particular, who have dedicated time to developing expertise in scientific disciplines other than statistics, such as postdocs from the successful GSP program, must have avenues other than traditional single-investigator grants for attracting outside funding.

It would be of enormous benefit to have NSF, preferably in conjuction with agencies having direct responsibility for the management of natural resources, to construct a long-term funding mechanism to support true interdisciplinary research among statisticians and geophysical, environmental, and ecological scientists. The initiatives undertaken over the past years have been highly successful in attracting statisticians to work in these sciences. There is now a need to foster the interdisciplinary aspects of that activity in a positive long-term manner.

One hopes that the NSF will continue to collaborate with the EPA, which has shown the initiative and funding potential to support large-scale projects in statistics, in particular through its support of two national centers for environmental statistics (at the University of Washington from 1996 to 2001, and at the University of Chicago from 2002).

However, with the exception of these two centers, EPA funding has been sporadic and uncoordinated. A further joint initiative by NSF and EPA could do much to attract more statisticians into the field, while at the same time the NSF's high standards of peer review and grant assessment

will ensure that such support is effectively targeted.

Other bodies with whom the NSF might productively collaborate include NOAA, NASA, DOE, NPS, and USGS. The first and last of these are particularly important for promoting statistical work of an ecological nature, NOAA in marine and USGS in terrestrial environments. From the perspective of academic researchers, as funding sources these agencies can benefit from cooperation with NSF in much the way that has been realized for the EPA.

5.4 Information Technology

The rapid rise of computing and large-scale data storage has impacted many human endeavors, sometimes in profound ways. There has never been a more exciting time for statisticians working in areas related to Information Technology (IT).

The development of the web and the exponentially increasing capabilities of computer systems have opened up undreamed of possibilities for the exchange of information, the ability to collect and analyze extremely large data sets of diverse nature from diverse sources, and to communicate the results. The development of open source software magnifies the ability of researchers to leverage their talents and ideas.

New challenges in statistical model building and learning from data abound. The efforts of statisticians and statistically trained scientists are having an important impact in all areas of science and technology, from astronomy to biology to climate to communications to engineering to intelligence, just to name a few at the beginning of the alphabet. Contacts with people in other scientific areas invariably give rise to opportunities for developing new ways to present, model and help interpret their experimental/observational/simulated data and to new methodology in experimental design and data collection.

The remainder of this section highlights a selected set of high-impact areas.

Communications

A wealth of communications records is generated every minute of every day. Each wireless and wireline call produces a record that reports who placed the call, who received the call, when and where it was placed, how long it lasted, and how it was paid for. Each user request to download a file from an Internet site is recorded in a log file. Each post to an online chat session in a public forum is recorded.

Such communications records are of interest to network engineers, who must design networks and develop new services, to sociologists, who are concerned with how people communicate and form social groups, to ser-

vice providers, who need to ferret out fraud as quickly as possible, and to law enforcement and security agencies looking for criminal and terrorist activities.

There is a host of challenging statistical problems that need to be met before the wealth of data is converted into a wealth of information. These include characterizing the probability distributions that describe the current behaviors of the millions of people generating the records, to updating the estimated behavior for each individual as the records fly by, to distinguishing the very small number of people with "interesting" behavior as early as possible. Perhaps surprisingly, these problems are inherently small sample, since most individuals do not generate huge numbers of records, and, not so surprisingly, are complicated by severe constraints on computing time and space. There is much that statisticians can contribute to solving these problems.

Machine Learning and Data Mining

The line between research in machine learning and data mining, as carried out primarily in Computer Sciences departments, and research in nonparametric estimation, as carried out primarily in Statistics Departments, has increasingly become blurred. In fact the labels 'machine learning' and 'data mining' are increasingly used by statisticians. Primary areas of very active research within Statistics Departments include new methods for classification, clustering, and predictive model building. Statisticians have been developing classification tools for a long time but the explosion in computational ability along with the fruits of recent research have led to some important new advances.

One such new advance in classification, taking advantage of these facts, is Support Vector Machines. The method is highly popular among computer sciences machine learning communities, but has greatly benefited from input by statisticians, who have contributed in important ways to understanding of the properties of the method. However there are important opportunities for further understanding of both the theoretical properties of this tool, and the most appropriate and efficient way to use this tool in recovering information from data, in a broad variety of contexts.

Recent applications of Support Vector Machines include: classification of microarray gene chips according to the type of disease carried by the subject donating mRNA to the chip, and classification of satellite radiance profiles according to whether and what kinds of clouds are present.

Examples of nonparametric risk factor modeling include joint risk of various medical outcomes, as a complex function of many risk factors. At a more exploratory level, clustering of mRNA signals by mixture modeling assists researchers to understand the number and nature of subpopulations in the data.

With the advent of high speed computing Statisticians are better able to build and test more sophisticated and detailed models that can deal

in a more realistic and interpretable way, with very large data sets and many potential predictor or attribute variables. It is important to put these models on firm theoretical and computational foundations to guide the applications.

Networks

The study of internet traffic can be roughly divided into traffic measurement and modelling, network topology, and network tomography. All of these areas present large scale statistical challenges.

Further research in measurement and modelling is motivated by the need to jointly improve quality of service and efficiency. The current approach to quality of service is based on massive overprovisioning of resources, which is both wasteful, and also not completely effective because of bursts in the traffic caused partly by improper protocol and routing procedures. Because many ideas for addressing these problems have been proposed, there is a major need for comparison, now done typically by simulation. This requires modelling, and seriously addressing the statistical problem of goodness of fit.

In particular, the central question is, "How do we know this works like the real traffic?" These issues present a host of new challenges to statisticians and probabilists. Classical statistical approaches and techniques are typically rendered impractical by the appearance at many points of heavy tailed distributions (often leaving even such standard tools as variance and correlation useless) and long range dependence and non-stationarity (stepping beyond the most basic assumptions of classical time series). However, understanding and modelling variation is still of critical importance, so this area is a large fertile ground for the development of creative new statistical methodologies.

Network topology presents different types of statistical problems. Here the goal is to understand the connectivity structure of the internet. Graph theoretic notions, combined with variation over time, and also combined with sampling issues, are needed for serious headway in this area.

Network tomography is about inferring structure of the internet, based only on the behavior of signals sent through it. Proper understanding, analysis, and modelling of the complex uncertainties involved in this process are important to headway in this area.

Data Streams

Statistical analyses of large data sets are often performed in what is essentially batch mode. Such data sets may require years to collect and prepare, and the corresponding statistical analyses may extend over a similar period of time. However, just as there exists an important niche in computer programming dealing with real-time computing and control, a rapidly growing niche for statisticians exists for real-time data mining. Such situations arise, for example, in remote sensing where limited bandwidth between an orbiting satellite and its ground station precludes transmission of all raw data. A

second example is commercial web sites such as an airline reservation system where detailed keystroke sequence data leading to actual or abortive reservations is not saved.

Off-line statistical analyses of these streams or rivers of data are not possible as the raw data are simply not available. However, a statistical agent can be placed directly in the stream of data to detect and quantify results typical of modern data mining. The challenge is to create statistical tools that run in almost linear time, that is, to design tools that can run in parallel with the real-time stream of data.

For simple statistics such as sample moments, there are no difficulties. However, these tools must be able to adapt in real-time. Furthermore, data mining makes use of virtually every modern statistical tool (e.g., clustering algorithms, trees, logistic regression). Transforming and recasting the statistical toolbox into this new and fundamentally important setting will require imagination, cleverness, and collaboration with algorithmic experts in other areas of the mathematical sciences.

More

Statisticians have played and continue to play an important role in other areas of IT, e.g., Medical Imaging, Computer Vision, Computer Graphics, Speech and Handwriting Recognition, Customer and Transaction Analysis, Document Organization and Retrieval.

5.5 Physical Sciences

Historically, astronomy is one of the first and most important sources of inspiration for, and application of, statistical ideas. In the 18th century astronomers were using averages of a number of measurements of the same quantity made under identical conditions. This led, at the beginning of the 19th century, to the method of least squares.

Astronomy has expanded enormously both in the size and complexity of its data sets in recent years in order to estimate the Big Bang cosmological parameters from the anisotropic clustering of galaxies, the fluctuation spectrum of the cosmic microwave background radiation, etc. A host of other basic statistical problems arises from the Virtual Observatory, a federation of multi-terabyte multi-wavelength astronomical survey data bases.

Despite the common origins of statistics and astronomy, and our mutual interest in data analysis, only very recently have there been substantial collaborations between statisticians and astronomers. (One example of this type was presented in the core chapter.)

This longstanding gap between the fields of statistics and astronomy exemplifies a common pattern in the physical sciences. Statistics works by the efficient accrual of evidence from noisy individual information sources. In large part the historical spread of statistical methodology can be described

as "noisy fields first": vital statistics, economics, agriculture, education, psychology, medical science, genetics, and biology. The "hard sciences" earned their name from the almost perfect signal-to-noise ratios attainable in classical experimentation, so it is understandable that they have proved the most resistant to statistical methodology.

However, recent trends are softening the hard sciences, and so there is an increasing need for statistical principles and methods. Technology now enables bigger and more ambitious data-gathering projects such as those of the Sudbury neutrino observatory and the Wilkinson microwave anistropy probe. These projects must extract crucial nuggets of information from mountains of noisy data. (The signal-to-noise ratio at Sudbury is less than one in a million.) Unsurprisingly, statistical methods play a big, sometimes crucial role in these projects.

To illustrate the promising future role of statistics in the physical sciences, we offer three brief statistics-intensive examples, from particle physics, chemical spectroscopy, and astronomy.

Confidence Intervals in Particle Detection

The following situation arises in the search for elusive particles: a detector runs for a long period of time, recording x interesting events; a similar run with the elusive particles shielded out yields a "background" count of y events. What is an upper confidence limit for the true rate of the particles of interest? Statistical issues become particularly sensitive if y exceeds x, so that the unbiased rate estimate is actually negative. The question then is whether the upper confidence limit is sufficiently positive to encourage further detection efforts.

Even in its simplest form–actual situations can involve much more elaborate background corrections–this problem has attracted widespread interest in the physics community. A much-quoted reference is Feldman and Cousins' 1998 Physical Review D article (p. 3873-3889). Louis Lyons, professor of physics at Oxford, has organized a September 2003 conference at the Stanford Linear Accelerator Center devoted to statistical problems in particle physics, astrophysics, and cosmology (www-conf.slac.stanford.edu/phystat2002/).

Comparative Experiments in Chemical Spectroscopy

Richard Zare, of the Stanford chemistry faculty, has developed an advanced class of mass spectrometers, able to simultaneously time the flights of large volumes of massive particles. This permits comparisons between collections of particles obtained under different conditions, for example complex molecules grown in different chemical environments.

A typical spectrum consists of particle counts in binned units of time, perhaps 15,000 bins in a typical run. Comparing two such spectra, that is looking for bins with significant count differences between the two conditions, is an exercise in simultaneous hypothesis testing. With 15,000 bins, the simultaneity is massive. Statistical methodology originally developed for microarray analysis can be brought to bear on spectroscopy compar-

isons, but the relation between time bins is unlike that between genes, suggesting that new methodology will be needed.

Survival Analysis and Astronomy

In a spectacular example of parallel development, astronomy and biostatistics invented closely related theories for dealing with missing data, the field called "survival analysis" in the statistics literature. The reasons for the missingness were different: astronomers are earth-bound so they cannot observe events too dim or too far away, leading to data "truncation". Data "censoring" occurs in medical trials when subjects fail to record a key event, such as relapse or death, before the end of the trial. Lynden-Bell's method and the Kaplan-Meier estimate, the astronomy and statistics solutions to the missing data problem, are essentially the same.

Mutual awareness did not occur until the 1980's. An influential series of joint astronomy-statistics conferences organized at Penn State by Babu and Feigelson have led to collaborations and progress in the statistical analysis of astronomical data. For example, the extra-galactic origin of gamma-ray bursts was demonstrated via survival analysis before the bursts could be identified with specific physical sources.

5.6 Social and Economic Science

Introduction

There is a long history of productive collaboration between statistics and the social and economic sciences. It is reflected in the fact that these disciplines routinely require basic statistics courses of all their undergraduate and graduate students.

An important feature of statistical work in these fields is the difficulty of obtaining reliable measurements when they come from people rather than physical instruments. For instance, obtaining salary data from individuals can be difficult since the same person might inflate his or her salary to impress a peer while deflating it when reporting it to a representative of the IRS. As a second example, data on illegal drug usage is difficult to obtain for the purpose of evaluating policies of deterrence.

Those who do statistical work in these fields have become especially adept at obtaining measurements in these situations. For them the quality of the data is highly dependent on the measurement methods.

Statistics as a discipline has always striven to abstract the essence of a problem and develop statistical methods that apply to it in the most general way. This often means that methods developed for one application are equally useful in another. Proportional hazards models developed for the life sciences are now used by economists. Survival analysis methodology developed in a completely different context is now used in the social sciences to model such things as time to job loss.

Some of the possible statistical breakthroughs in the economic and social sciences may come not only from developing entirely new methodologies but also from adapting existing ones from other fields.

Areas of Opportunities for Statistics

Many research opportunities exist for statistics in the social, behavioral and economic sciences and there is need for true interdisciplinary collaboration. Economists and social scientists seek to develop and test models of the world and statistics has always been critical to that.

Computationally intensive methods have freed them from the straight jackets of models that have convenient mathematical properties but do not conform to theory or observed data. The emphasis is shifting from tractable asymptotic modeling to computationally described models, which are more interesting, theoretically coherent, and realistic.

The collaborations between these scientists and statisticians now requires that the statisticians become more grounded in the theory of the science discipline and that the discipline scientists be able to express their theories in general form.

The complexity of models has lead to an increase in the complexity of associated decision and risk analyses. The use of prior information and Bayesian methods has become important for creating and assessing the results, pointing to a need for new research in Bayesian statistical decision theory. An exciting example is the use of Bayesian hierarchical modeling and decision analysis for the choice of strategy in home radon measurement and remediation. It provided a new decision strategy that would save the same number of lives as the current recommended strategy for about 40% less cost.

One of the straight jackets of modeling has been independence and significant opportunities exist for modeling dependency across units and over time. (The assumption of independence takes the social out of social science. It fundamentally decouples the methods from the core of the theoretical process.) Examples of applications that would benefit from this include network modeling for complex processes, dynamic modeling, and experimental design for simulation studies and complex systems. Important advances have been made in analyzing categorical responses that exhibit clustering through the dependence introduced by repeated measurements for the same subjects over time or by measurements for families.

The importance of the models in these disciplines also cries out for more statistical attention to uncertainty in the model rather than just the sampling. The very question of how close data-based models can come to the "true" underlying model is a very important question and some of the limitations have been explored.

Sampling and survey methodology are basic tools for describing and evaluating the models of the social and economics sciences. The problems of bias or nonresponse in sampling and lack of precision due to small sample sizes have been ameliorated through the development of more flexible Bayesian

models that take into account the features of the survey design. Even in situations where the data measurements for complex physical models do not come through people, the development of complex models themselves may require obtaining opinions, estimates and subjective probabilities from people in many different disciplines (chemists, physicists,engineers, biologists) to build the model. Once again the special statistical tools of measuring data from people are effective.

Some outstanding opportunities exist for interdisciplinary interactions in modeling complex systems for human populations and their economic and environmental interaction through computer simulations. The opportunities range from parameter-based computer experiments to agent-based simulation, i.e. simulation-based models where multiple entities sense and stochastically respond to conditions in their local environments. These computer generated responses allow us to observe the evolution of social and economic processes.

Critical statistical methodologies include the use of statistical experimental design to effectively choose values of input for the simulations. Furthermore, the presence of prior probabilities about portions of the models make the use of Bayesian analysis (often hierarchical) particularly effective. The models themselves are profitably scrutinized through sensitivity and uncertainty analyses of input and output. The complex and often massive data output from the simulations can profitably be searched for patterns and meaning through the use of statistical high dimensional data reduction techniques and multilevel analyses.

Simulation modeling examples include alternative policies and interventions for land use change and ecological effects at the rural urban interface. They include the social and economic organizations of lake users to test design manipulations that remove invasive species of fish from a lake. They include the dynamic landscape of land use and land cover change in Thailand. They include the emergence of cooperation among human groups based on human-environmental interactions.

6

Statistical Education

6.1 Overview

Clearly, the long-range solution to the shortage of statisticians must lie in improvements to the educational system that will attract, train, retain, and reward a new generation of talented students. Improvements will be required across the spectrum from elementary school through continuing education of the workforce. The pool of K-16 teachers who are qualified to teach statistics needs to be increased.

As a result of the stunning growth in AP statistics and other efforts to incorporate college level material into the high school curriculum, the boundary between high school and college statistics has blurred. The typical college introductory statistics course often does not push far enough in terms of fulfilling the needs of undergraduates for skills in data analysis, modeling, and experimental design. These and other pressures suggest that the entire K-16 approach to statistical training needs to be reevaluated to assure that there is a logical and complete progression. At the same time, the sharing of best practices needs to be emphasized to speed change and increase efficiency.

According to CBMS2000, undergraduate enrollments in statistics courses in mathematics and statistics departments were up sharply—45%—between 1990 and 2000. Faculty growth has not kept pace, and statistics departments, like mathematics departments, appear to be relying increasingly on temporary faculty, according to this report.

While it appears that graduate doctoral training has become more balanced in recent years, it is under stress from an enlarged statistics core, increased demand for cross-disciplinary work, and expansion of the intersection of statistics and computer science. The situation is delicate and will require departments to adapt in response to their local challenges and opportunities. Post-doctoral training should receive higher priority in the statistics community as a sensible way to launch research careers and broaden interests. NSF's VIGRE activity is providing a significant boost in this regard.

Statisticians should play a larger role in the science of educational reform whether it involves statistics instruction per se or other topics. There is a cross-the-board national need for decisions about education to be based on sound data!

So much of what needs to be done to strengthen education in statistics will require large investment that it is natural to look to NSF for leader-

ship and support. The shortage of statisticians in key industries and many government agencies, the heightened concerns about national security, and the increased reliance at the college level on adjuncts and instructors who are not members of the regular faculty all suggest that now is the time for action.

6.2 K-12 and statistics

There has been a general movement towards more emphasis on flexible problem solving skills, including working with data, throughout the K-12 system. While not without controversy, this trend seems to be a healthy one from the perspective of increasing the statistical literacy of the population.

Now with AP statistics courses, a significant number of students are seeing statistics before they get to college. It may be too soon to assess the impact of this development on the pipeline for future statisticians. But it does seem clear that this full-year course, taught in an interactive high school environment (in contrast to a one semester college lecture format), has a lot going for it.

The role of the teacher cannot be over emphasized in all of these developments. Considerable effort continues to be given to training K-12 teachers to take on AP statistics. Still, the need exceeds the supply. Dealing with the K-12 (and college level) shortage of trained teachers may well be the number one priority in statistics education. There is a special opportunity and challenge for assuring that a qualified corps of teachers is available in the inner cities and other places with large minority populations.

Generally, the number of minority students interested in technical studies has been small. This can be traced in part to limited exposure in middle and high schools. Improving statistics literacy is obviously only one leg of a larger challenge in mathematics and science preparation.

6.3 Undergraduate Statistics Training

With the demand for undergraduate statistics courses on the rise, as mentioned above, and the growing appreciation of the value of a statistically literate society, it is only natural to expect increased emphasis on undergraduate statistics training. Special pressure points can be seen in the demand from the pharmaceutical, biomedical, and government statistical communities (including those involved in homeland security and national defense matters). The demand includes both undergraduate and graduate levels of preparation.

The introductory statistics course—" stat 101" —has received considerable attention in recent years. Certainly, much progress has been made

in the development of course content, pedagogy, and technology and how they relate to each other. However, the progress is hardly complete (and never will be) and must continue in order to reflect the ongoing evolution of the subject. Another concern is the " trickle down" effect that can result in out-of-date training being provided at community colleges and other places where there is too long a time lag for course updates.

There are two other issues with respect to stat 101. First, with the advances in K-12 training, many students are entering college with exposure to statistics that may even surpass the ordinary introductory course. Hence, there is the need to rethink matters from a holistic K-16 perspective. Second, stat 101 may not go far enough to meet the needs of many undergraduates. The leap into a mathematical statistics course or into a specialized course within a major is not the answer. There is a growing recognition of the need for a natural sequel, a stat 102, that will help students continue to learn about data analysis, modeling, statistical computing, and experimental design—practical skills that will prepare them for success in particular application domains.

Going one step further, a minor-in-statistics option may be attractive for undergraduates who require extensive skills of this kind. It is a logical alternative that falls between the limited exposure of a stat 101-102 series and a full-blown statistics major. It may be the right answer for many social science and some engineering majors.

Engineers are often faced with the need for data collection, experimentation, and decision making in the face of uncertainty. All of these involve basic statistical concepts and methods. Yet, with the exception of industrial engineering, statistics is not a part of the core curriculum in most (if not all) major engineering programs. At the same time, industry has a critical need for sound statistical thinking among its engineers (and others). For example, it spends enormous amounts re-educating its workforce in basic statistical methods for quality improvement. NSF could help close this gap in undergraduate engineering training by bringing together educators from engineering and statistics and industrial leaders to figure out how to address this problem.

Concern over the state of undergraduate education in statistics is hardly new. Bryce (2002) points out that it has been there for over 100 years. What has changed, he observes, is that there is now a critical mass of statisticians and educators attempting to improve matters. This couldn't be happening at a better time.

6.4 Graduate Statistics Training

Statistics doctoral training has improved considerably over the past decade in the sense that there is a better balance between theory, methodology

(including data analysis and computing), and probability. However, the increasing breadth of the field is causing difficulties. Questions are being asked as to what constitutes the core of statistics and what is essential to teach graduate students. Some worry that decreased emphasis on mathematics will lower the ability of graduates to tackle and solve truly difficult and complex problems. Others argue for more diversity in statistics graduate programs, recognizing the impossible task of dealing effectively with an expanding core of " required" knowledge without watering down the curriculum too much. It is critical to get this balance right.

In addition to the expansion of the field and the fuzziness at where its edges lie, there is also the growing intersection of statistics and parts of computer science. Quite what this means for statisticians and graduate education is still being debated (Breiman 2001). Nevertheless, a strong case can be made that today's typical graduate program reflects an overly narrow, pre-computing view of statistics. Much of the statistics that emerged from this era was centered on problems of estimation and testing where mathematics brilliantly finessed a paucity of computing power. While acknowledging the continuing demand for this type of classically trained statistician, there is increasing concern that students emerging from graduate programs are not adequately prepared to engage in cutting-edge research and collaboration in the newer information-rich arenas. The skills needed go much deeper than just the ability to program or to use statistical packages. They include, for example, the ability to design and evaluate algorithms for computationally challenging statistical methodology.

With many younger statistics faculty immersed in cross-disciplinary research, it is natural to expect that graduate students will increasingly latch onto research topics coming from these exciting application areas (e.g., genomics, neural science, finance, and astronomy). A natural next step would be the formal development of cross-disciplinary doctoral programs involving statistics. At the very least, experiments along these lines should be encouraged and monitored with the goal of finding models that work.

A related recommendation comes from the Odom report. It proposes that graduate (as well as undergraduate) mathematical sciences education be broadened to include " areas other than mathematics." It goes on to suggest stronger support for programs that involve " mathematical scientists in multidisciplinary and university/industry research."

Another opportunity is to invest resources in developing graduate courses in statistics tailored to the needs of particular disciplines such as atmospheric sciences. Besides providing a needed service, an important side benefit could be the sparking of new inter-disciplinary collaborations.

Statistics, along with virtually every science, lags in the production of minority doctorates. The current percentage of under represented minorities among the total of statistics doctorates appears to be about five. Yet there are models available on how to run a successful program. Ingredients include attractive fellowships, a critical mass of minority students, and in-

dividualized attention from faculty.

In most universities, a professional master's degree in statistics is at least a significant portion of the graduate program. The basic course work usually places less emphasis on mathematics, theory, and specialized advanced courses, and more on methods, consulting, and computing. Employers often still feel that such training needs to be stiffened in terms of applications, data analysis, and communications skills. Nevertheless, these degrees are an important part of the equation for increasing the pipeline of statisticians—especially those who are U. S. citizens.

An intriguing variation is a master's program aimed at doctoral students in other disciplines. This is the natural extension of the undergraduate minor idea mentioned in Section 5.3. It is also an alternative to the cross-disciplinary Ph.D. degree.

In contrast with mathematics education, there is no well-established discipline called statistics education. This lack of formal structure and study is an impediment to the development of new ideas and their incorporation into educational practice. To overcome this situation, some universities are beginning to establish degrees in statistics education. The American Statistical Association is also focusing on this issue through a working group on undergraduate education.

Given all of these pressing concerns, tensions, and opportunities surrounding graduate training in statistics, it would be especially timely and beneficial to organize a series of national workshops to explore how best to deal with these challenges.

6.5 Post-Graduate Statistics Training

Post-doctoral appointments have never been a major component of the career path for statisticians, in contrast to many other disciplines. One reason for this has been a long-standing very strong job market. Graduates can find immediate employment in career-path positions without such experience. So why bother?

Notwithstanding this reality, carefully crafted post-doctoral appointments can be an immensely rewarding career-enhancing step. Their availability and use should probably be expanded. The Odom report argues for graduates in the mathematical sciences to use these opportunities to immerse themselves in other disciplines. The National Institute of Statistical Sciences (NISS) has provided 43 government and industry-sponsored post-doctoral appointments over the past decade. Consistent with the mission of NISS, the primary emphasis has been on cross-disciplinary training. It is anticipated that the new Statistics and Applied Mathematics Institute (SAMSI) will offer a substantial number of post-doctoral appointments in the years ahead. Even for graduates who desire a traditional research path

within the core of statistics, these appointments can offer rich opportunities for maturation, increased professional breadth, and experience in securing research funds without the pressure of a tenure clock.

Other types of post-graduate training can be equally valuable for career enhancement. These include mentoring on the job by seasoned professionals, early sabbaticals, and a whole range of formal continuing education programs aimed at keeping statisticians at all levels up to date with new developments.

6.6 Special Initiatives (VIGRE)

NSF continues to sponsor a variety of initiatives of interest to the statistics community. The Grants for Vertical Integration of Research and Education in the Mathematical Sciences (VIGRE) activity is especially relevant to the transformations that many statistics departments desire to make. Still, the program has left many departments with the impression that it is mainly appropriate for mathematics departments. Statistics departments that lack an undergraduate major or are relatively small in size feel at a disadvantage, for example.

Nevertheless, several departments of statistics have been beneficiaries of VIGRE awards and reported that their impact has been very positive in several ways. This is not hard to understand because the goals of VIGRE, e.g., integration of research and education, enhancing interactions, and broadening experiences, are very natural ones for statisticians.

For these departments, VIGRE has resulted in strong undergraduate research experiences and increased the flow and improved the quality of students entering graduate school. It has also impacted how graduate departments think about training. Post-doctoral appointments in statistics have increased as departments have been able to make appointments that otherwise would have been infeasible.

Questions remain as to how these positive changes will be sustained. Clearly, this will boil down to money and faculty commitment to a large extent. Nevertheless, most of the experiences reported have been enthusiastic. The challenge for the statistics community and NSF will be to capitalize on what has been learned and to spread the knowledge beyond the institutions that have been part of the VIGRE activity. An NSF-sponsored conference may be an effective way of doing this.

It would be worth considering other types of programs that could broaden the experiences of students such as strengthening ties between university departments and government and industrial groups.

6.7 Continuing Education

For statistics graduates of all kinds, as well as novices to the field coming from other backgrounds and disciplines, there is a very clear-cut need for continuing education. The demand for such training is well illustrated by the large audiences (including professors) at tutorial sessions at the Joint Statistical Meetings and the expansion of the Continuing Education Program of the American Statistical Association. Much of this demand comes from industry and government where there are the unfortunate countervailing pressures of the need to learn while on the job and the constraints of lack of time and money for such training.

While it is now well understood that educational technology is only part of the solution for K-16 and graduate training, there should be a much larger role here for technology-intensive solutions. By taking advantage of the capabilities of the Web and exploiting various multi-media technologies, one can envisage a range of distance learning experiences that are practical and cost effective for the mature student with limited time and budget.

6.8 Educational Research

For obvious reasons, it is essential that educational reform programs dealing with statistics (or any other subject for that matter!) be based upon sound statistical studies of their benefits. Yet there is a sense that too many decisions about education are based on anecdotal information. Statisticians should be particularly sensitive to these risks. They are also in a position to assist education researchers in beefing up their studies so that the next time someone asks for the data it will be there.

7

Summarizing Key Issues

Evidence is all about us for the current unique opportunities for Statistics. The three pillars of the Mathematical Sciences Priority Area of the National Science Foundation, for example, consist of handling massive data, modeling complex systems and dealing with uncertainty. All three are core activities of the discipline of Statistics.

Massive amounts of data are collected nowadays in many scientific fields. But without proper data collection plans this will almost inevitably lead to massive amounts of useless data. Without scientifically justified methods and efficient tools for the collection, exploration and analysis of data sets, regardless of their size, we will fail to learn more about the often complex and poorly or partly understood processes that yield the data.

In order to master this enormous opportunity, the profession must address several important challenges. Some of these are the intellectual challenges that we have reviewed in the preceding chapters. A second set of challenges arise from the external forces that face the profession. In this chapter these external challenges will be recapitulated.

7.1 Developing Professional Recognition

Earlier in this report some time was spent on identifying "What is statistics" and on reviewing the history of the profession. The reason for this is simple.

The role of the statistics profession is often only poorly understood by the rest of the scientific community. Much of the intellectual excitement of the core of the subject comes from the development and use of sophisticated mathematical and computational tools, and so falls beyond the ken of all but a few scientists.

Statistics is the discipline concerned with the study of variability, with the study of uncertainty and with the study of decision-making in the face of uncertainty. As these are issues that are crucial throughout the sciences and engineering, statistics is an inherently interdisciplinary science. Even though it does not have its own concrete scientific domain (like rocks, clouds, stars, or DNA),it is united through a common body of knowledge and a common intellectual heritage.

Statistics is no longer, if it ever was, just another mathematical area like topology, but rather it is a large scale user of mathematical and computational tools with a focused scientific agenda. The growth of the profession

in the past twenty years is enormous. For example, the number of doctoral degrees granted in statistics has grown steadily to the point of matching the number of degrees in the "rest" of mathematics.

If we wish our separate needs to be met, we need to establish our identity to the wider scientific audience. We hope that this report can contribute to this goal.

7.2 Building and maintaining the core activities

Exploitation of the current manifold opportunities in science has lead to an increased demand for greater subject matter knowledge and greater specialization in applications.

This in turn has created a challenge for statistics by putting the core of the subject under stresses that could with time diminish its current effectiveness in the consolidation of statistical knowledge and its transfer back to the scientific frontiers. In essence, the multidisciplinary activities are becoming sufficiently large and diverse that they threaten professional cohesiveness.

If there is exponential growth in data collected and in the need for data analysis, why is core research relevant? It is because unifying ideas can tame this growth, and the core area of statistics is the one place where these ideas can happen and be communicated throughout science. That is, promoting core area statistics is actually an important infrastructure goal for science from the point of view of efficient organization and communication of advances in data analysis.

A healthy core of statistics (through a lively connection with applications) is the best hope for efficient assimilation, development and portability between domains of the explosion of data analytic methods that is occurring. As such, it is a key infrastructure for science generally.

In Chapter 4 we identified the following opportunities and needs for the core:

- **Adapting to data analysis outside the core**. The growth in data needs provides a distinct challenge for statisticians to provide, in adequate time, intellectual structure for the many data analytic methods being developed in other arenas.

- **Fragmentation of core research**. Outreach activity is high and increasing for all sorts of good reasons. We think there has been an unintended consequence of this growth – a relative neglect of basic research, and an attendant danger of our field fragmenting.

- **Manpower problem.** There is an ever shrinking set of research workers in the U.S. who work in core area research. This manpower

problem is destined to grow worse, partly from the general shortage of recruits into statistics and partly because outreach areas are pulling statisticians away from core research.

- **Increased professional demands**. The core research of statistics is multidisciplinary in its tools: it borrows from (at least) information theory, computer science, and physics as well as from probability and traditional math areas. As statisticians have become more and more data-focused (in the sense of solving real problems of modern size and scope), the math skills needed in core areas have gone up. This need for ever increasing technical skills provides a challenge to keeping the core vital as a place for integration of statistical ideas.

- **Research funding.** It seems clear that funding for core research has not kept pace with the growth of the subject. Investigators, rather than beating their heads against difficult funding walls, turn their efforts towards better funded outreach activities or consulting. The most basic needs remain as they always have: to encourage talent, giving senior people time and space to think, and encouraging junior people to buy into this line of research.

- **New funding paths.** An illustration was given of a possible funding route that might enable statisticians to enrich basic statistical research with interdisciplinary activity without pulling them out of core research.

7.3 Enhancing collaborative activities

A distinguishing feature of the intellectual organization of statistics is the value placed on individual participation both in the development of statistical methodology and on multidisciplinary activities, e.g., in applications of statistics in biology, medicine, social science, astronomy, engineering, government policy, and national security, which, in turn, becomes an important stimulus for new methodology. Although different people strike different balances between methodological research and subject matter applications, and the same people strike different balances at different times in their careers, essentially all statisticians participate in both activities.

Statistics, through these interactions, develops tools that are critical for enabling discoveries in other sciences and engineering. Statisticians are also instrumental in unearthing commonalities between seemingly unrelated problems in different disciplines, thereby contributing to or creating synergistic interactions between different scientific fields.

However, as noted by the Odom Report, our reach has not been wide or far enough:

> Both in applications and in multidisciplinary projects, however, there exist serious problems in the misuse of statistical models and in the quality of education of scientists, engineers, social scientists, and other users of statistical methods. As observations generate more data, it will be essential to resolve this problem, perhaps by routinely including statisticians on research teams.

One problem is that statisticians who attempt to participate widely in these activities face several steep challenges, including the need to stay current in all relevant fields and the need to provide the relevant software to carry out statistical analyses. In addition, in spite of a culture that encourages multidisciplinary activities, evaluating these activities has proven difficult and sometimes controversial.

From Chapter 6, Statistics in Science and Industry, we can identify the following important issues:

- The large amounts of data produced by modern biological experiments and the variability in human response to medical intervention produce an increasing demand for statisticians who can communicate with biologists and devise new methods to guide experimental design and biological data analysis.

- There exists a software challenge that touches deeply in a number of areas. It corresponds to a wide need for statistical methods to be incorporated into open source software products, and a corresponding lack of support for this infrastructure need.

- There exists a need for coordinated long term funding for interdisciplinary projects so that the statistician can afford to develop the scientific understanding vital to true collaboration.

7.4 Education

We have identified a growing need for statistics and statisticians from wide areas of science and industry. As noted by the Odom Report, "There is ample professional opportunity for young people in statistics, both in academia, industry, and government." At the same time, the profession cannot meet demand with domestic candidates. Again, from the Odom Report: " A very high proportion of graduate students are foreign-born and many remain in the United States upon graduation."

At the same time that the demands on the profession have grown in the research arena, there has been a startling growth in demand in statistical education at lower levels:

- The statistics profession has started to feel the impact of the wide growth of statistical training in grades K-12, led by the implementation of an AP course in statistics. This means many students are coming to college with previously unheard of knowledge about statistics.

- Undergraduate enrollments in statistics are up sharply–45%–between 1990 and 2000.

These circumstances point to the need for the profession as a whole to consider how to handle this growth, and how to build a statistics education infrastructure that can meet the changing and growing needs. Here are some of the key issues and needs:

- The need for trained teachers of the Statistics AP courses, as well as statistically literate instructors in other subjects in grades K-12.

- The need for an integrated K-16 curriculum that accounts for the better high school training.

- The need for expanded statistics minor and major options in both undergraduate and graduate programs.

- The need to encourage and enable students to acquire deeper and broader subject matter knowledge in an area or areas of application.

- At the graduate level, there is a large challenge in building training programs that can offer sufficient depth over the wide breadth of tools that the modern statistician is currently using.

- There is a growing need for more postdoctoral training opportunities to help newly minted graduates develop their professional skills.

A second set of challenges to the statistics profession comes from the need to fill the pipeline for tomorrow. The number of people with training in statistics is not growing nearly fast enough for the exponential growth in the demands for statistical expertise. This trend must be changed dramatically in order to meet the high demands for statistical expertise for today's complex, large, interdisciplinary research problems in science and engineering.

No doubt recruitment at the lowest levels has been helped by the AP course. At the same time programs for enhanced recruitment that are focused on the mathematics profession as a whole, such as VIGRE, are very promising, but have many times lacked sensitivity to the special needs of statistics.

8

Recommendations

The statistics profession faces many challenges at this time. The scientific program of the workshop was very helpful as a way of identifying the broad needs of the profession. In this report we have tried to summarize the key elements of the workshop talks and presentations. In the end, this leads us to ask if there are recommendations we can make to the statistics community and to its constituents that will direct attention in the right directions. The Odom report provided the following very useful summary of its recommendations to NSF regarding mathematics as a whole:

> Therefore, NSF's broad objective in mathematics should be to build and maintain the mathematical sciences in the United States at the leading edge of the mathematical sciences, and to strongly encourage it to be an active and effective collaborator with other disciplines and with industry. NSF should also ensure the production of mathematical students sufficient in number, quality, and breadth to meet the nations's needs in teaching, in research in the mathematical sciences and in other disciplines, and in industry, commerce, and government.

We wholeheartedly second this recommendation. Here we would like to focus more specifically on those areas of opportunity and need in the statistics profession. The scientific committee therefore recommends that (1) the profession adopt the following long term goals and (2) that the profession ask its key stakeholders, including universities and funding agencies, to support these goals.

Promote recognition of the unique identity of statistics. This involves clarifying its role vis à vis the other sciences. This should include making clear the difference in roles and needs between statistics and the other mathematical areas. The committee hopes its report will be one step in this direction.

Strengthen the core research areas. This requires ensuring that their full value in consolidation of knowledge is recognized and rewarded. Achieving this goal would be aided by expansion of the current funding programs in this area.

Strengthen multidisciplinary research activities. This can be accomplished by helping all parties involved calculate the full cost of multidisciplinary research and so make allowances for its special needs. This would imply research support that recognizes the special role of statisticians in scientific applications, where they are toolmakers (concepts, models, and

software) as well as tool providers.

Develop new models for statistical education. This should be done at every level, and separately for majors and non-majors. Attaining this goal will require the focused efforts of experts in statistical education working together to create and disseminate the models.

Accelerate the recruitment of the next generation. For this to succeed, the profession needs the cooperation of the key stakeholders in developing programs that recognize the special needs and opportunities for statistics.

Appendix A

The Workshop Program

This appendix contains the program that was given to the workshop participants. (Some changes were made to the program during the workshop in order to accommodate the schedules of the Foundation participants.)

STATISTICS: CHALLENGES & OPPORTUNITIES FOR THE 21st CENTURY

In recent years, technological advances in instrumentation development and the exponential growth of computing power have allowed researchers to collect large amounts of data. Examples include the data collected from the Hubble telescope or satellite photometry in the physical sciences, genetics and the data bases it has spawned in the life sciences, and internet-related data in engineering and the social sciences. Common characteristics of all these data sets are size, complexity and noisiness. These massive data sets and their collection create new challenges and opportunities for statisticians, whose vital role is to collect data, analyze it and extract information from it.

At the same time, science, industry and society are now tackling a multitude of inherently data-poor situations, such as subsurface pollution monitoring and remediation, reliability of complex systems, such as nuclear plants and materials, and study of vehicle crash-worthiness. This is being done by a combination of mathematical/computer modeling and statistical analysis, and requires optimal use of scarce, and hence invaluable, data. This poses new challenges and opportunities for statisticians who must optimally design experiments in situations of extreme complexity and then extract the maximal information from the limited data.

Thus, whether it is because of the multitude of new data-rich or of new complex, data-poor situations, it is a critical time to assess the current status and needs of the field of Statistics to ensure that it is positioned to meet these challenges. In this context, it is important to address the following questions:

- What is Statistics?
- What are the distinct features that define Statistics as a discipline?
- Given that, over the last 50 years, many major universities have separated Statistics and Mathematics, what are the characteristics that distinguish Statistics from Mathematics?
- Given that Statistics is in the Division of Mathematical Sciences (as part of the Statistics and Probability program), what is its appropriate share of funding for the mathematical sciences? What tools, data or resources that are needed to address this question, now and in the future?

- What are the current and future exciting research directions and opportunities in Statistics?
- What are the interactions of Statistics with other disciplines? In particular, what does Statistics contribute to these disciplines? And how do these disciplines benefit from Statistics?
- Tools from Mathematics are used in Statistics, and vice versa. But is there enough collaborative research involving both Statistics and Mathematics? If not, what is preventing this from taking place?
- What are areas of scientific investigation in which statisticians should be involved but are currently not?
- What should be the goals for the discipline of Statistics over the next two decades?
- What is needed, for example in terms of human resources and facilities, to achieve these goals?
- What is the role of Statistics in the international scene?
- What are the funding trends?
- How do other disciplines perceive Statistics and Statisticians?
- How do the statisticians perceive themselves?

Remark: The overall purpose of the workshop is to examine exciting research directions and the discipline of Statistics.

PARTICIPANTS:

Scientific Organization Committee:

*Jim Berger, Duke University
*Peter Bickel, UC Berkeley
Mary Ellen Bock, Purdue University
Lawrence Brown, University of Pennsylvania
Sam Hedayat, University of Illinois at Chicago
Bruce Lindsay, **Chair**, Pennsylvania State University
David Siegmund, Stanford University
Grace Wahba, University of Wisconsin
* indicates person is also a speaker

Speakers:

Sir David R. Cox, Oxford University, UK
Iain Johnstone, Stanford University, IMS President
Jon Kettenring, Telcordia
Vijayan Nair, University of Michigan
Eric Feigelson, The Pennsylvania State University
Chris Heyde, Australian National University, Australia, and Columbia University
Joel Horowitz, Northwestern University
Werner Stuetzle, University of Washington
Warren Ewens, University of Pennsylvania
Richard Smith, University of North Carolina, Chapel Hill
Philippe Tondeur, Division Director, NSF/DMS
Robert Eisenstein, Assistant Director, NSF/MPS
Adriaan De Graaf, Executive Officer, NSF/MPS
Rita Colwell, Director, NSF
Joe Bordogna, Deputy Director, NSF

Other Confirmed Participants:

Roger Koenker, University of Illinois Urbana-Champaign
Martina Morris, University of Washington
Alan Agresti, University of Florida
Wing Wong, Harvard University
Bruce Levin, Columbia University
Michael Stein, University of Chicago
Peter Guttorp, University of Washington
Karen Kafadar, University of Colorado-Denver
Jeff Wu, University of Michigan
Alan Karr, NISS
Regina Liu, Rutgers University
William Padgett, University of South Carolina
Peter Hall, Australian National University, Australia
Willem van Zwet, Eurandom, The Netherlands
Nancy Reid, University of Toronto, Canada

Keith Worsley, McGill University, Canada
Robert Tibshirani, Stanford University
Brani Vidakovic, Georgia Tech
Mitchell Gail, NIH
Steve Marron, University of North Carolina Chapel Hill
Gary McDonald, General Motors Co.
Augustine Kong, deCODE Genetics, Iceland
David Scott, Rice University
David Madigan, Rutgers University
Stanley Wasserman, University of Illinois Urbana-Champaign
William B. Smith, American Statistical Association, Executive Director
Miron Straf, National Academy of Sciences, ASA President
Mark Kaiser, Iowa State University
Robert Kass, Carnegie Mellon
Diane Lambert, Bell Labs

SCHEDULE

Day 1: May 6, 2002, Room 375

8:30- 8:45 Breakfast

8:45- 9:15 Rita Colwell, Director of the NSF

9:15- 9:30 Robert Eisenstein, Assistant Director, MPS

9:35-10:25 What is Statistics? (D. R. Cox)

10:25-10:35 Break

10:35-11:25 The core of Statistics (Johnstone)

11:25-12:15 Statistics and the Biological Sciences (Ewens)

12:15- 1:00 Lunch (**Lunch Speaker:** Dr. Joe Bordogna, Deputy Director, NSF)

1:30- 2.20 Statistics and the Geophysical and Environmental Sciences (R. Smith)

2:20- 3:10 Statistics and the Social and Economic Sciences (Horowitz)

3:10- 3:25 Break

3:25- 4:15 Engineering and Industrial Statistics (Nair)

4:15- 5:05 Statistics and Information Technology (Stuetzle)

5:30 Program committee meets to discuss Day 2.

7:00 Dinner

Day 2: May 7, 2002, Room 110

8:30- 9:00 Institutes: The role and contribution to Statistics (Jim Berger)

9:00- 9:30 Statistics in the International Scene (Chris Heyde)

9:30-11:30 Funding in the Statistical Sciences: Current modes and future models

9:30-10:00 NSF perspective (Philippe Tondeur, Division Director, DMS)

10:00-11:00 Panel Discussion

Panelists: Feigelson, Heyde, Tondeur, Berger

Moderator: Bruce Lindsay

Floor Discussion

11:30-12:30 Lunch (**Lunch Speaker:** Adriaan De Graaf, Executive Officer, NSF/MPS)

12:30-5:30 Separate Groups are formed (discussion and writing. Please refer to the end of this program for information about the composition of the different groups and room numbers.)

5:30 Program Committee meets to discuss Day 3

DAY 3: May 8, 2002, Room 110

8:30-9:50 All groups join for a general discussion. The group moderators will present the group's summary to the entire body for discussion. The resulting (revised) summaries will form the basis of sections for the final reports and other publications.

9:50-11:00 Break (room not available)
11:00-12:15 Continuation of discussion and presentation
12:15-1:15 Lunch
1:15-2:15 Vision 2020 (Kettenring and Bickel)

Discussion Groups:

The composition of the groups is as follows (with members of the organization committee serving as moderators).

SOCIAL AND ECONOMIC: *Moderator*: Mary Ellen Bock
Participants: Agresti, Morris, Horowitz, Koenker, Wasserman, Straf
BIOLOGICAL SCIENCES: *Moderator*: David Siegmund
Participants: Wong, Levin, Kong, Gail, Worsley, Tibshirani, Ewens, Kass
INFORMATION TECHNOLOGY: *Moderator*: Grace Wahba
Participants: Marron, Stuetzle, Madigan, Scott, Lambert, Feigelson
ENGINEERING AND INDUSTRIAL: *Moderator*: Sam Hedayat
Participants: Kafadar, Nair, Liu, Kettenring, Padgett, McDonald, Karr
GEO AND ENVIRONMENTAL: *Moderator*: Larry Brown
Participants: Guttorp, Smith, Vidakovic, Stein, Kaiser
CORE: *Moderators*: Bruce Lindsay and Jim Berger
Participants: Johnstone, Cox, Hall, Reid, Heyde, Van Zwet, Wu, Bickel