American Statistical Association

Promoting the Practice and Profession of Statistics

732 North Washington Street, Alexandria, VA 22314 USA
(703) 684-1221 ● Fax: (703) 683-2307 ● *president@amstat.org*
www.amstat.org

February 28, 2013

Iain Johnstone and Fred Roberts
Co-Chairs, StatsNSF
Mathematical and Physical Sciences Advisory Committee
The National Science Foundation
4201 Wilson Boulevard
Arlington, Virginia 22230

Dear Professors Johnstone and Roberts,

I write with the American Statistical Association's (ASA) response to your calls for comment on how the National Science Foundation (NSF) can better support the statistical sciences. This communication includes both an ASA recommendation and a selection of comments from our members.

We appreciate the work of your committee on this enormously important issue. Statistical science is the science of collecting, analyzing, and understanding data, as well as accounting for the relevant uncertainties. As such, it permeates the sciences and is intertwined in all facets of the NSF's work. The field of statistics is ideally situated to translate methodological advances from one application area to another, thereby benefiting all areas of science, and doing so in an economically prudent way.

We believe NSF would benefit greatly by substantial and agency-wide measures to better support the statistical sciences. Indeed, with NSF having been created "to promote the progress of science," better recognition of and support for the statistical sciences at NSF would better position the foundation to carry out its mission.

The ASA's recommendation is that a chief statistical scientist be added to the office of the NSF director. The purpose of this position would be to leverage advances in the statistical sciences across all NSF divisions to support the agency's mission of advancing sciences as rapidly and economically as possible. The responsibilities would include the following:

i) Interacting with all directorates on statistical sciences (i.e., what statistics is, the limitations of data, the design of experiments, the analysis of data, the interpretation of the analysis, the limitation of the methods, etc.)
ii) Strengthening every application area of science by promotion of interdisciplinary research (through the interaction of scientists who specialize in data with those who specialize in the application area)
iii) Leveraging investments (and advancements) in the statistical sciences by coordinating among directorates
iv) Regular reporting on opportunities, successes, and funding of statistical sciences
v) Promoting the statistical sciences across the federal government and within the NSF
vi) Chairing a funded intra-agency working group of program officers from all the directorates tasked with creating cross-cutting programs involving statistics

We considered the placement of this new position at a level other than the office of the NSF director, but are convinced any lower placement would undermine the position's ability to effectively interact across the directorates and the federal government. To respond to those who might suggest the DMS director could meet these responsibilities, we stress the importance of this being a new, full-time position—separate from existing operational responsibilities—in the NSF director's office to adequately address challenges and opportunities. A division director must necessarily be focused on the science of that division, and the goal of the chief statistical scientist is broader, to make the best choices for science independent of a particular division.

We understand an additional position in the NSF director's office runs contrary to recent efforts to streamline it. Nevertheless, we fervently believe the discussions to date point to the need for a major change at NSF to address the challenges of better supporting the statistical sciences and this new position is the best and most viable solution. We also believe our recommendation aligns with the OneNSF goals of enabling seamless operations across organizational and disciplinary boundaries, empowering the NSF to respond to new challenges and leveraging resources and opportunities for maximum impact.

The ASA issued two calls for comments in response to StatsNSF requests. Our first call was in November, when ASA leadership was asked for comment about how NSF could better support the statistical sciences. The second call was in January, when ASA leadership received from you five questions to pose to its membership.

A selection of responses can be found in enclosures A and B. We now summarize those comments, starting with the responses to our first call. There were many good comments from ASA members, but none have the substantive and agency-wide impact of a chief statistical scientist in the director's office.

A common—but not new—suggestion is a separate division of statistics, with interdisciplinary programs combining statistics and all areas of science. While this suggestion has many merits, it is not clear where administratively one would place such a division to effectively interact with all directorates or whether a new division is viable at this time.

Other alternatives are to have statistical program officers in all directorates overseeing interdisciplinary programs that support new collaborations between statisticians and other research areas or to fund institution-based statistical science centers to provide support to NSF funded projects.

ASA members suggested other ideas that would not have agency-wide impact yet have significant merit. An idea that stands out in this category and that the ASA endorses is to expand upon the 11 programs currently supported by DMS so the statistical sciences are represented by more than one. The current structure, in which statistics is just one of 11 programs, feeds the misperception that statistics is a narrow discipline with little overlap to other scientific fields. Among the additional programs that should be strongly considered are computational statistics, biostatistics, and environmental statistics.

For both calls for comment, a common theme was the importance of recognizing statistics as an independent, mature *scientific* field. One responder put it this way:

> One thing that has become clear to me in my years working within a mathematics department is that mathematicians prefer a precise answer to an approximate question, whereas statisticians prefer an approximate answer to a precise question. That is, if someone poses a real-world question, a mathematician will replace it with a somewhat similar question that can be precisely answered, while a statistician will come up with a useable, if only approximate, answer to the question that was actually asked. That's an important difference.

Reviewing the responses to our call for comment, we received many interesting ideas and wish we had the resources to pursue an in-depth follow-up or exploration of them. Instead, we propose these questions be continually pursued as a responsibility of the chief statistical scientist. A wide selection of comments received from the second call is included in Enclosure B. We do not summarize them all in this letter, but following are a few common themes.

A prominent theme to emerge was that the statistical sciences need to be part of research in many areas, not just as “statistical support,” but true integration of statistical sciences into the domain sciences so methodological approaches can be developed that address the specific challenges posed by the scientific area. The current NSF structure does not adequately support this type of activity.

The responses to questions 3–6, which probed the scientific areas that might benefit from more engagement of the statistical sciences, consistently made a strong case for statistical sciences having a bright future with broad applicability across science and benefitting science.

As a final note, several ASA commenters felt NSF DMS wouldn’t fund statistics that isn’t mathematically elegant enough or is more applied or computational, suggesting there may be a perception problem among statisticians regarding their chances of being funded by the NSF.

These are exciting times for science, with many new challenges and opportunities presented by the deluge of data from recent inventions, social media, and high-throughput genomic

technologies. The ASA strongly believes NSF will be better positioned to tackle these challenges and opportunities with a chief statistical scientist guiding the integration of the statistical sciences with other science disciplines at the nation's premier scientific agency.

In closing, the ASA's leadership thanks you for your work on this committee.

Sincerely,

Marie Davidian, PhD
President, American Statistical Association

Enclosures:
Enclosure A: Summary of ideas received from the ASA's first call for comments
Enclosure B: Selection of comments received from the ASA's second call for comments

Enclosure A:

Summary of selected comments received by the ASA in response to our first call for comments to StatsNSF call for comments on the general charge for how to better support the statistical sciences at NSF.

1. Separate Division of Statistics with program officers having expertise across all areas of science
2. Create positions statistics program officers in all Directorates who are associated with a program and who would have their own proposals and pot of money
3. The funding of reproducibility studies mainly consisting of statistical re-analyses of data from experiments with important findings
4. Create at least one disciplinary new program within the Division of Mathematical Sciences, focusing on areas that fall within the Statistical Sciences, but have a strong standing of their own
5. Greater support from NSF for preparing those who teach statistics
6. Mechanism for developing relationship with statistics program officers. In our community, program officers are sometimes seen as distant arbiters rather than partners, fundamental for the success of their careers.
7. More continuity through review process (so resubmissions build on previous submission and it isn't a start-over).
8. Collaboration in Mathematical Geosciences-like opportunities for statisticians: interdisciplinary funding opportunities requiring a statistician
9. Make outreach to statistical community part of program officer job description so they don't have to shoe-horn in such outreach
10. More permanent statistics program officers (without taking away from rotating number)
11. "I wish that I had specific proposals to present, but I don't. I can only say that I would be better off, my college would be better off, and society would be better off if statistics were more clearly separated from mathematics in the mind of the general public. Too often students who "don't like mathematics" turn away from statistics because they think that it is a branch of mathematics. Others study statistics but tend to miss important (non-mathematical) ideas because they are focused on the mathematical aspects of statistics.

    "One thing that has become clear to me in my years working within a mathematics department is that mathematicians prefer a precise answer to an approximate question whereas statisticians prefer an approximate answer to a precise question. That is, if someone poses a real-world question, a mathematician will replace it with a somewhat similar question that can be precisely answered, while a statistician will come up with a useable, if only approximate, answer to the question that was actually asked. That's an important difference. Pushing statistics under the tent of mathematics impedes social progress."

Enclosure B:

Selected comments received by the ASA in response to our second call for comments to address the specific questions from StatsNSF. The comments are grouped into General Comments and Question-Specific Response. Each number under each is from a different person.

The six questions are

1. What should NSF do to further promote and facilitate the appropriate development of statistical science? Are there management structures that should be considered?
2. Is research support in statistical science not requested from NSF because it lacks a home? If so, what might be a possible remedy?
3. Are there complex or massive data problems that might be amenable to joint attack by several disciplines? If so, please specify.
4. What are some examples of disciplinary areas that could benefit from statistical science methodologies that are already being employed in other areas?
5. What are some examples of simultaneous development of statistical science methods for different fields that might benefit from cross-fertilization?
6. Are there research areas in statistical science that, with sufficient funding support, could spur significant advances in science? If so, please specify.

General Comments

1. I was Director of Statistics and Probability at NSF …. While I was there, I conducted numerous discussions amongst various programs at NSF regarding the establishment of a Division of Statistical Sciences. Rather than Statistics be perceived as one of various mathematics programs within the Division of Mathematical Sciences, I felt that there were a sufficient number of statistics-related programs within NSF that could be plucked from their current locations and relocated within a Division of Statistical Sciences. The discussions were very fruitful and I thought that it was an idea whose time had arrived.

   The Directors of the disciplines of mathematics within the Division of Mathematical Sciences were very unhappy with such talk of a separate Division for Statistics because of possible repercussions that a move such as this would have back on them…

   I believe that such a move has to happen and I applaud the Statistical Societies in their quest to get something done to improve the visibility of Statistics within NSF and the scientific community.

2. I am concerned about the recent (what I perceive to be a) sharp decline in the number of American applicants for graduate work in Statistics and Biostatistics. It seems as though every year, the number of undergraduate students with adequate preparation declines--is this a problem nationally? The NSF Graduate Research Fellowships target the top students. Can we cast a wider net? Of course, [Summer Institute in Biostatistics] is meant to address this in part, but, still, I'm concerned that soon we will just not have enough qualified students to fill the surging demand.

3. Thank you for requesting member input regarding NSF and the statistical sciences. I would like to see an increase in the cross communication among statisticians and disciplines requiring analysis and modeling of large and complex systems. In particular, consider the following two areas of considerable activity.

   From Wikipedia (I know, but forgive me) we find the following definitions:

   > **Statistical physics** is the branch of physics that uses methods of probability theory and statistics, and particularly the mathematical tools for dealing with large populations and approximations, in solving physical problems.
   >
   > **Data mining** is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

   Wouldn't it be fun to have these groups interacting in a manner similar to that forwarded by the Santa Fe Institute? Statistical physicists from the sciences trying to talk to data miners from business, medical and national security areas would surely stimulate new approaches for dealing with complex modeling problems.

4. Thank you for the opportunity to respond to your request. Let me begin by stating that I am a social work professor with good quantitative skills (there are actually a good number of us). As such, we have a strong person-in-environment perspective upon which our discipline is based.

   Therefore, we are looking at integrating and strengthening statistical techniques such as spatial analysis and social network analysis. Often we work in interdisciplinary teams that may include public health, urban planning, medicine, psychology and education. So below is my input.

5. Not sure what question this answers, but to me the main problem with the present arrangement is that it is very difficult for a statistician who cannot get NIH money to raise the funds to have a real group, including funding for postdocs. Statisticians who do nothing but prove theorems can be funded like mathematicians, but statisticians who work in interdisciplinary areas or in algorithm development need to be able support real research groups. Furthermore, new PhDs in these areas should all be doing postdocs before starting faculty positions. I'm fairly agnostic as to the right bureaucratic structure to achieve appropriate funding levels to make this possible.

6. A few thoughts.
   There are lots of areas with massive data that really require joint ventures among different types of quantitative scientists (statisticians, applied mathematicians, signal processors, computer scientists, data miners, bioinformaticians)

   Here are a few:

   1. Application areas yielding quantitative image data (especially in neuroimaging). Clearly statisticians need to be heavily involved to sift through all the many sources of variability and correlation and multiple testing in the modeling, but computer

scientists/image processing people are needed for many of the processing and segmentation tasks that are complex, essential, and not necessarily very statistical.
2. In high dimensional genomics data, again statisticians are needed because of the many sources of variability, multiple testing issues, and integration across data type problems, yet the data are enormous so data management is key plus this area has lots of compiled information on the web about genes, proteins, etc. that are needed for modeling so bioinformaticians are needed to write programs to pull this information and integrate with the raw data for analysis.
3. Another example is astronomy where the data are huge and computer scientists and signal processors can effectively work with statisticians to build models to pull insights from these data. In fact, you see statisticians getting into signal processing areas (e.g. compressive sensing) to work on these problems.

Also, there are many application areas where cross-fertilization of statistical modeling and methodological ideas are sorely needed. I think a big reason this does not happen is that there are not strong incentives for statisticians developing methodology to branch out to many applied areas and try to port these ideas over across different fields. This can be most effectively done by master applied statisticians, but unfortunately there may be a gap of information between the newest methodologies and what is known by applied statisticians. If we can improve this information flow of innovative methods to different fields of application clearly our profession can increase its impact.

7. Because statistical support can facilitate research in so many disciplines, one model for the NSF to consider would be to fund statistical consulting units directly to provide support to other NSF supported projects as well as possibly other projects at universities. One problem with the model where only specific research projects are individually supported is that this approach to funding research works against universities operating as communities with free flow of ideas and more spontaneous collaboration formation. Instead, collaboration and the flow of ideas between disciplines has increasingly been linked to the flow of $$'s which can stifle collaboration and risk-taking. Perhaps the NSF should consider providing support to statistical consulting units on university campuses for the purpose of fostering collaboration more generally and identifying problems that have statistical solutions.

8. Having recently earned my PhD in Survey Methodology at the University of Michigan, with a focus on the statistical aspects of survey methodology, I feel a bit surprised that there are not more outlets for support of original research in survey statistics. Since I graduated in 2011, I've submitted four proposals to the Methodology, Measurement, and Statistics (MMS) section of NSF (three this January), and there seems to be a widespread belief among my peers that there are simply no other outlets where survey methodologists can submit proposals at NSF.

   I feel that there are a large number of areas that require important methodological and statistical research in survey statistics, but very few funding options within the federal system. Attempts to have these kinds of projects funded by NIH will always require some kind of health context or topic, which is not ideal for survey methodology in general. And I get the sense that submitting proposals to MMS results in our proposal being evaluated by economists and other theoretical statisticians who have little background in the issues that are most important for survey statistics.

In sum, it would be nice if there was a section at NSF specifically dedicated to funding original research in survey methodology. AAPOR and ASA have recently collaborated to introduce a new journal entitled the *Journal of Survey Statistics and Methodology* dedicated specifically to publishing this kind of research. Now we just need more options for having this research supported!

9. Have you considered posing these set of questions upon a larger audience? The scope is rather limited. Not knowing if she is a member, I did forward this to Hilary Mason.

10. I believe that it is important to ensure that statisticians are not viewed as just data analysts, but as scientists that are involved from the onset of any research project. I also believe that statisticians, by participating in multidisciplinary research teams, will be able to contribute suggestions for the identification of measurable responses, for appropriate data collection plans, for best data analysis methods, and correct interpretation of results. As a result of participating in these research teams, new research in statistics will result. Henceforth, I suggest there be a small group of statisticians within the NSF that educates researcher in other fields about the scientific role of statisticians and identify statisticians within academic or government institutions that can be assigned to support different projects.

11. The $ 200 million 'Big Data' initiative launched by the Obama administration is aimed to transform their ability to use huge data generated at the global level for scientific discovery, environmental and bio-medical research, education, and national security. The role of statistics in this initiative is of paramount importance needing a far more important structure and support than at present. I compliment ASA under your leadership to take a lead on this issue.
    The success of the Initiative is heavily dependent on statistical methodology and computational approaches. 'Data science' as a blend of statistical, mathematical and computational sciences is emerging as a new discipline to this end. Way back in 1920s we had 'small' data sets to handle, leading to a great flowering of statistical methods like Student's 't' and various inference methods for small samples. We are now at the other extreme needing methodological development and applications to inter-disciplinary fields as indicated in the set of questions raised in your e-mail.

12. I am a member of both SIAM and ASA and have read the communications from the presidents of both societies with great interest. Although the request from SIAM has slightly different language than the request from ASA, using the term "data science" wherever the ASA communication uses "statistical science", I am writing the same response to both requests. I live and work in the Washington, DC area and have been involved with both statistic and mathematical communities through research, consulting, teaching at the undergraduate and graduate level, and several tours as program director at the National Science Foundation.
    In my view, the current developments in data science (used broadly, including data collection in various disciplines, computational treatment, statistical analysis, mathematical modeling, and communication and visualization) represent an abundance of scientific opportunities that I have not seen in three decades of professional activity. Here is an opportunity to find new challenging and relevant research problems for all, to attract the best student talent to a field that is suddenly very "hot", and perhaps to bridge some of the gaps which have arisen, in my view often unnecessarily, between mathematics, statistics, and computer science. Let's not waste this opportunity.

13. Very briefly, in my opinion statistical science is in danger of being overtaken by computationalists who don't know enough about inference. So to me, broad statistical areas in need of support are study design, causal inference, and adjustment for missing data, and sensitivity analysis. I'm not sure these areas have been supported very much for large scale data analysis. More specific areas that I think need strong support are the science of measurement (metrology) and informatics. Philosophically, I also think statisticians need to rethink the value of hypothesis testing. It conditions on unobservable quantities (parameters) rather than observed quantities (the data), which I think confuses our field to no end.

14. By way of introduction and establishing my vantage point: I am a Professor and former Chair (2005-12) of the Department of Statistics at the University of South Carolina. I am an ASA Fellow and President 2011-12 of the Southern Regional Council on Statistics. I am an applied statistician, with about half of my 70-odd publications outside of mainstream statistics journals. I have had regular federal funding most of my career (before I became department chair), mostly as a statistical consultant in large multi-investigator ecological research projects funded by NSF, NOAA and others.

Statistics already plays a central role in most sciences; as more powerful computers appear on our desks and in our hands, as terabyte data-captures become commonplace, it will only grow in importance. Yet development of statistical theory and methods is perennially beholden to others. Ecologists (many of them my dear friends, by the way) will support the development of statistical tools as long as these will clearly be of benefit to ecologists in the near future. One can't blame them for that – every dollar given to a statistician is a dollar that could have gone to an ecologist. Replace "ecologist" with psychologist, geographer, medical researcher, economist, chemist, biologist… our hat is always in our hand, at every port. Who funds long-term foundational statistical ideas? Who nurtures this infant science, barely a hundred years old, for its own sake?

The reflex answer offered by most is mathematics. Formal statistical inference is based on the mathematical theory of probability, so the Division of Mathematical Sciences at NSF seems the natural home for statistics. Mathematics and statistics are in fact nearly opposites, though. Statistics is about dealing with uncertainty – formalized inductive reasoning based on data. Mathematics is about being certain – there are few arguments more final, more certain, than a deductive mathematical proof. As a former member of a mathematics department (some of them my dear friends, by the way) I believe that accepting uncertainty, embracing it as a fact of life – is abhorrent to most mathematicians. Brilliant as they are, most do not even understand the difference between probability and statistics. The DMS has been anything but a nurturing home for statistics. Most statisticians are discouraged workers who gave up on DMS long ago because their ideas were not mathematically elegant enough to win grants there. It just wasn't good mathematics.

The fact that formal statistics is built on the mathematical theory of probability doesn't mean statistics should be considered a subfield of mathematics. All of chemistry follows from physical principles at the molecular and atomic levels – yet chemistry is not a subfield of physics. We would not leave funding decisions for the core development of chemistry in the hands of physicists – the result would look a lot more like physics than like chemistry.

15. We write this letter in response to the request transmitted by the American Statistical Association (ASA) for comments about NSF support of statistical sciences, to be transmitted to the StatsNSF Committee.

    We write from our perspective as officers of the Statistical Ecology section of the Ecological Society of America. We strongly believe that increased funding targeted to increase the flow of ideas in both directions between statistical researchers and the application domains of ecology, evolution, and organismal biology would generate great advances. Although there are many excellent statistical scientists working in these research fields, the population of such people is small relative to the magnitude and diversity of important data problems they generate. [This letter has not been officially approved by the Section or Society due to the short turnaround required.]

16. I offer here one fundamental change for the organization of statistics research at NSF.

    The change is based on my vision of statistics as infrastructure for interdisciplinary research, in which statisticians work together with scientists in other disciplines to increase their powers of observation; to further their abilities of measurement, analysis, and prediction; and to advance their discoveries and guide their responsible development into technologies or other societal benefits that make a difference in people's lives.

    The change for NSF to help achieve this vision is one I advanced in my speech as ASA President in 2002:

    > The National Science Foundation (NSF) should establish interdisciplinary programs in its research directorates that focus on data, measurement, and the development of statistical methods to advance their sciences. A model is the Methodology, Measurement, and Statistics Program in the Directorate for Social, Behavioral and Economic Sciences. Most statistics research at the NSF is managed through a program in the mathematical sciences, just like mathematical subdisciplines, as number theory and analysis. But statistics is no longer just a branch of mathematics. Computer and information technologies are replacing mathematics as the foundation for more and more statistics research. At the very least, the NSF should establish a division for statistics research that is separate from mathematics.

Responses to Question 1: What should NSF do to further promote and facilitate the appropriate development of statistical science? Are there management structures that should be considered?

1. Add several disciplinary research programs within the field of statistics. Statistics has only one program while Mathematics has 10 (including probability), a clear indication of imbalance. Could aim for Applied Statistics, Computational Statistics, Nonparametric Statistics...

2. I have been in the business industry for 19 years. The term Statistics is often misunderstood, and the field Statistics is often undervalued. However, the demand for statistics type professionals has grown significantly over the past many years, due to initiatives such as Business Analytics, Machine Learning, Big Data, and Data Sciences. Most companies (other than in pharm and biotech) that apply heavy statistics do not call their stat professionals statisticians – rather, they tend to be called data scientists, quantitative analysts, business analysts, etc.

3. There needs to be a systematized guidance for incoming and incumbent staff on the proper incorporation and use of statistics. If there is a board or a leadership team, there needs to be an Amstat member (or equivalent) represented at the same hierarchical level as the rest of the members as an indication to NSF that statistics are as integral as the other components.

4. I think the Statistics program is small in NSF with limited funding resources. I would be glad to see more interactions, connections, or cross and joint funding opportunities of the Statistics program with other divisions (or programs) in NSF, not only restricted to DMS.

5. The DMS/NIGMS program is an outstanding place to highlight new statistical methods and their application in the biological sciences. This is a good outlet for statisticians whose work is often "at the border" between NSF DMS and NIH. I hope this program is continued.

6. I have the impression that NSF funds only very theoretical statistical research. I served on an NSF review panel some 15 years ago, to pre-screen proposals (Statistics and Probability, I think). It seemed that there was no interest (except for mine) in the handful of proposals aimed at applied, methodological research. That might have changed, but I still hear from colleagues that you are wasting your time submitting a proposal to NSF for methodological research.

7. I believe that there are management structures that are essential for there to be useful and beneficial statistical science that supports research in the social and physical sciences. Every management structure constituted as a research team should include a mathematical statistician with advanced training and a certain minimal amount of applications experience (i.e., a senior professional). This should be adopted as a core requirement for sound research and it should be promoted extensively as a principled way of doing research by and for NSF. Decision-making in a modern society is driven by data and data-analysis. Statistical professionals are too often brought into the data analysis stage of research project work long after data have been collected and asked to provide answers to research questions either when essential data-elements have not been quantified or when hypotheses have been posed after the data have been collected. Sometimes a consequence of the former is that data for some variables required for the application of a statistical technique have not been collected; in the latter case, it may not be possible to satisfactorily address important research issues because no provision was made for collecting the necessary data.

8. One thing that might help is to combine machine learning, data mining, and statistics in some ways. ML and statistics have different cultures, but they are interested in many of the same topics, and some people feel that they are the same actually (or at least a large overlap). I feel that both fields would benefit from some interaction.

   Also I believe that NSF should encourage projects that are *translational*, in other words, projects that can translate into something that will truly help science or society. I discuss this more below.

   Another thing I realized is that NSF should do more to promote women and minorities to the greatest extent possible. In machine learning, there are very few senior women at the top

universities. It is very difficult to name more than a few top women in ML. Even at the women in machine learning workshop at NIPS, the senior speakers are not all that senior. Our field has a serious problem with women not becoming top academics (my guess is that it is not by choice), and I can't even fathom to begin how to fix it. But I think NSF could really help with the right kind of encouragement.

9. Communicate and collaborate with other NSF divisions relative to emerging issues in science methodology where mathematics and statistics can contribute.

   NSF tends to give credit in grant awards for proposals that claim prior results with a p-value of <= 0.05. Such claims are not consistently evaluated by statistically competent reviewers. Making such reviews a regular part of the NSF culture generally would do more to raise the awareness of and respect for statistics than any other single step.

10. Perhaps NSF's most important contribution in "further promoting and facilitating the appropriate development of statistical science /data science" could be in graduate education. This is a very exciting field for students. There is also a well identified talent gap. Create and support opportunities where students can learn applied mathematics, statistics, and computing (the emphasis will necessarily fall on the last two fields), with a view towards data science, and the best researchers in these fields will work on topics that can attract these students. The current management structures at NSF allow for supporting statistical and mathematical work from within DMS, but computing is mainly supported in a separate directorate that has a very different culture. A solution cannot just be found within MPS (the directorate of mathematical and physical sciences).

11. Balancing the role of exploratory theoretical work with the need for applied statistics is an important issue; of course these two perspectives can interact as well.

12. We believe that support for applied methodological research in fields such as ecology, evolution, and organismal biology is a thin area of NSF funding programs. It is possible to obtain funding, but it is a difficult “needle to thread”: typically proposals must water down the statistical advances to communicate well to biology panelists or, on the other hand, must be sold to statistical panelists who may not see applied work in different application domains as valuable.

    We think a funding mechanism with the following two features would be extremely valuable for fields such as ecology, evolution, and organismal biology:
    1. Explicit support of applied methodological research.
    2. Support of new collaborations between statisticians and biologists.

    By “applied methodological research”, we include the following:
    1. Research that translates or adapts recent statistical innovations to specific application domains.
    2. Research that builds theoretical foundations for the methods that emerge rough-hewn from application domains.
    3. Development of new methods for specific application domains for which the initial research outlets will be biological journals.

17. Review committees should include statisticians. Their comments generally make other reviewers aware of issues that have been overlooked – the end result being some level of education and appreciation.

Responses to Question 2: Is research support in statistical science not requested from NSF because it lacks a home? If so, what might be a possible remedy?

1. To certain degree, yes. The current Statistics program seems to emphasize on Mathematical Statistics or theoretical Statistics. For researchers working on biostatistics, computational statistics, etc., it may be difficult to get good reviews. Again, more connections with other NSF divisions would be helpful.

2. No, it's likely because statistics, as with mathematics in a broader sense, is looked at as a tool for sciences, not a science unto itself. Statistics shouldn't try to carve out its own research niche, but instead be a required component of good science generated by NSF.

3. I believe statistical science needs to maintain a home in NSF especially because it cuts across so many disciplines.

4. Is there a home at NSF for statistical methodological research? I'm unaware of one.

5. My colleagues *do* request research support from the NSF, and I am supported on an NSF CAREER award. The model for how research is done in statistics and CS is culturally different. Many statistics departments won't admit additional graduate students even if faculty members have support for those students for 4 years. It is believed that this model is more stable somehow to trends and availability of funding. In computer science (where ML usually sits) we always support our students and postdocs ourselves, through NSF support or in industry. I admit that I like that model much better, because it allows me control over the research I would like to participate in. If I can get funding for the project, I can do it. Definitely if there were a clear home for statistics in NSF it would be easier to find funding opportunities and apply for competitive grants.

6. Its home is supposed to be in the Division of Mathematical Sciences, which should request more proposals for development/expansion/innovation of statistical science. Its current web site and awards list barely mention it and provide only one award out of 60 listed.

7. I think an exclusive statistical science department would be beneficial. Like in most major universities, statistical science has its own department. The benefits could include, but are not limited to, the following:
    a. Visibility; it is not under umbrella of other departments
    b. Convenience of collaboration among statisticians; chat during lunch hours, stop by and do some scratches on the board without a formally scheduling a meeting
    c. More efficient and concerted management of efforts in different research area; avoid duplicated effort

d. If funding goes directly to the department of statistical science, the funding could be prioritized. Though, the downside could be less funding for those programs deemed less important.

Of note, all the justifications for an exclusive statistics department in a university could be applied here.

8. I cannot answer the question whether "research support in the statistical sciences / data science is often not requested from NSF because it lacks a home there". I note that the culture of support in statistics is somewhat different, with statisticians often obtaining partial research support from agencies other than NSF (EPA, USGS, NIH, to name a few). With respect to data science, the field is too young to even make such a statement, since, well, there seem to be very few data. So I regard this as a somewhat loaded question.

9. Yes. We suggest two possible remedies:
    1) Weave into existing funding programs such as those in the BIO directorate the possibility of special collaborative proposals on statistical methodology. These could be inspired by past programs such as QEIB, which were more broadly oriented toward "quantitative" or "mathematical biology" projects and emphasized theoretical biology more than statistical methodology.
    2) Initiate a new program specifically for advancing statistical methodology and practice in any application domain. Such a program could emphasize or require collaborations between statisticians and biologists (or other scientists). Such a program would be distinct from the Advances in Biological Informatics (ABI) program, for example, by casting a wider net of statistical applications than the "innovation/development" format of ABI. The latter is outstanding it its own goals, but precludes many types of proposals that could advance statistical methods and practices.

Responses to Question 3: Are there complex or massive data problems that might be amenable to joint attack by several disciplines? If so, please specify.

1. Too many of them – Big Data has become a popular term which includes sources from social media, web analytics, macroeconomics, genomics, astronomy, quantitative finance, etc.

2. Yes, certainly any physical or computer sciences with computer horsepower can team with statistics to work on big data. Climate change, social media marketing, JPL NEO asteroid tracking, etc.

3. Statistical research on genomic data may be considered by both DMS and the Biology division. Statistical research on computational issues, e.g. database security, online and instant responses, could be considered by both DMS and Computer Science.

4. Yes, there are certainly complex and big data with which we must deal. In particular many of us in social work are beginning to use agent-based modeling to better understand individual preferences and how this leads to the emergence of various phenomena. We also deal with

massive record that have little utility at the actual program management level. Again we have a large interest in applied spatial analysis and network analysis.

5. Yes, this is true in many areas. In medical imaging, there could be better ways to measure improvement or progression of disease by using statistical analyses at the pixel level. In genetics, the analyses of large microarray data requires careful statistical considerations. The scientists that are trying to tackle these problems using numerical analysis techniques and high computing power are often not aware of the consequences that lie within the data dependencies that exist.

   I also think that statistical science can contribute to all disciplines that collect and analyze data via research areas in experimental design.

6. Most definitely. The IT people are needed for efficient data capture/storage/analysis techniques, but are often clueless on issues of variance, probability, estimation, and the vast field of simulations that underlie currently accepted statistical techniques. Both IT and statisticians are often clueless as to what underlying physical structures are likely or possible. One can see IT proceed with zero interest in causal structures. Both groups are often clueless as to what should be the next step in the field's development. Thus, three groups are needed: subject specialists, IT, and statisticians. If the research is anticipated to impact population behavior/political decisions/economic investment then, at the planning stage, are needed social psychologists and sociologists/political scientists (or politicians)/economists (or investment experts). Both implementation of what is learned and what the next steps need to be part of the initial research effort.

   One sees overly optimistic claims made by some computer-intensive techniques – tree pruning, cube slicing, etc., which greater scrutiny from the statistical perspective would have guided towards more realistic claims.

7. Yes, certainly. I think industry could also be involved, as there are many companies now that work on big data.

   One major problem is that, currently, the ML and statistics fields are so focused on theory and methodology, that it is difficult to get *truly* applied papers published. We have gotten to the point where most real-world problems can be solved with existing methodology, and the difficult part is in the application, and really understanding the problem in order to develop a knowledge discovery system. However, the field is still focused mainly on theory and methodology, and papers that apply the techniques to a new application area (even a critical area like energy or medicine) without a methodological component, will not get accepted to an ML conference, ML journal, or statistics journal. Even in the Annals of Applied Statistics, they expect methodological content.

   I have been trying to remedy this to some extent, where I am co-editing a special issue of the Machine Learning Journal that is called "Machine Learning for Science and Society". It is focused on how ML is currently helping science or society. One problem we noticed is that authors often contribute work that will never be deployed, because there is no incentive to deploy it, as one cannot publish this in an ML journal. So very few people projects ever come to fruition to help science and society! This is a serious problem we need to remedy. The goal should go

beyond just publishing the paper; it should be to actually help people. Domain experts from whichever discipline or industry the work is in should be involved.

The NIH has been moving towards funding more translational research, and I believe the NSF should do this as well. To do this, there should be some evidence of impact to society, and deployment.

8. The Big Data Initiative of 2012 calls for an interdisciplinary approach to addressing complex data structures. Sampling strategies and the use of simulation need to be integrated with analysis of complex and massive data problems.

   The future of computation will lie in quantum computing; work should be proceeding now to take advantage of emerging opportunities for computational statisticians in this area.

9. Yes, absolutely. All these massive data problems come from real applications. To understand large data sets from astronomy / climate science / marketing, the mathematical scientist needs to be a member of a team that has astronomers / climate scientists / marketing experts. The mathematical scientist may have to give up the intellectual driver's seat for such projects and should accept this. The boundaries between very high-level consulting and original research may be vague and undefined.

10. Clearly, the biotech area will provide statisticians with a full plate.

11. Yes, ecology and environmental science in general are generating increasingly massive data sets both from long-term monitoring programs and from technological advances in field data collection. Examples include:
    a. Continent-scale long-term monitoring data such as the Breeding Bird Surveys of North America and some European countries.
    b. Synthesis datasets such as those created by working groups at the National Center for Ecological Analysis and Synthesis. Examples of synthesis datasets can be found in the Knowledge Network for Biocomplexity database, DataONE, or at NSF-funded Long-term Ecological Research (LTER) sites.
    c. Future data streams from the National Ecological Observatory Network (NEON).
    d. Continental-scale ecosystem gas flux data from NSF-funded networks such as FLUXNET.
    e. High-resolution animal movement and behavior data from advanced marking and telemetry technology.
    f. Use of machine learning and internet crowd-sourcing for problems such as arthropod population counts, animal identification (by species or individuals) from camera traps, and land-use change.
    g. US Forest Service Forest Inventory Analysis (FIA) data
    h. Citizen Science data such as eBird, The Great Sunflower Project, OakMapper, iNaturalist, and others.
    i. And, of course, the increasing application of genetics and genomics in ecology, such as in the developing field of landscape genetics.

12. Certainly in medicine there are several examples (although they would be directed toward NIH): mining electronic health records joins clinicians, informaticists, and statisticians; determining

genomic signatures joins biologists, clinicians, and statisticians; assessing the impact and accounting for missing data is a general problem that would require a content expert as well as a statistician; likewise causal inference.

Responses to Question 4: What are some examples of disciplinary areas that could benefit from statistical science methodologies that are already being employed in other areas?

1. The question is what are the examples that cannot benefit from statistical science methodologies – I can think of possibly English literature and not much more.

   Typical examples that could benefit from statistical science include:

   1) Clinical trials – the hottest area for applied statisticians
   2) Genomic analysis
   3) Causal inference as applied to epidemiology, social sciences, and business – the focus on causality in the statistics field has not been sufficient, leading to the knowledge growth of causal inference in other fields such as AI, epidemiology, economics, and philosophy. One can combine knowledge from these disciplines to support causal inference.
   4) Marketing and Business Analytics – this is an area that is seriously under-represented in ASA documents and discussion, so it is being taken by other quantitative fields but formally trained applied statisticians continue to make significant contributions; the career opportunities in this area are currently huge and only growing hugely given growing data size and increasing focus on data-driven decisions in business; it also has much similarity with personalized medicine
   5) Risk management – another area seriously under-represented, but with the economic change in 08, firms have been hiring more statisticians and econometricians in this area
   6) Quantitative finance – also somewhat under-noticed, so the field has been taken over mostly by physicists who mostly apply statistical methods
   7) Insurance – more insurance firms are hiring statisticians or stat type professionals to personalize premium rate
   8) Economic and business forecasting
   9) Supply chain management and logistics
   10) Strategic finance
   11) Astronomy – heard there are tons of data coming down from the sky every night but there're limited resources to analyze them statistically

   In summary, the majority of industries in our world are business in nature (for-profit or non-profit) but statistics is seriously underrepresented in the business world despite the large and growing demand there. As a result, the business world is seeking analytics professionals by hiring data scientists, marketing analysts, decision scientists, marketing scientists, risk analysts, etc., and the job nature is quite often statistical. It has been observed that once employers discover the efficiency and effectiveness of a formally trained applied statistician, they tend to hire more and more of them.

2. A disciplinary are that could benefit from statistical science methodologies is program evaluation. I'm an external evaluator on NSF and U.S. DoEd grants that aim to increase the recruitment and retention of STEM undergraduates. I am also pursuing a Ph.D. in educational research and evaluation, with an interest in measurement. In reviewing the STEM evaluation literature, I see a

lack of psychometrically tested instruments to measure attitudes toward STEM fields, although increases in attitudes is a common objective in these grants.

Thus, there is no way to assess across programs the extent to which they are increasing STEM attitudes other than retention and graduation rates, and to determine which approaches may be more effective than others. Also, I'm not sure as to the extent to which many evaluators can help their clients develop educational measurement instruments for curriculum redesign projects. Funding for the development of statistical and measurement skills among evaluators is needed for NSF, DoEd and other agencies to truly meet their own mandate for scientifically rigorous evaluations.

3. The interdisciplinary work involving the intersection of geography, urban planning, public health and behavioral economics have the most promise for us.

4. Every subject that generates data would benefit from statistical science. Sociology, behavioral, nutrition, medicine, anthropology, business, sensory science, chemistry, microbiology, botany, and all physical sciences. Now, the hot trend is predictive analytics and all these master degree programs.

5. One sees journals with titles of XX Statistics, where XX is Medical, Sociological, etc. Such journals often explain techniques long presented in statistical journals, as opposed to presenting new techniques. One conclusion is that the articles in statistical journals are not providing the level of understanding needed for a subject-area expert to implement the technique. XX Statistics journals do present the technique in terms used by XX practitioners, with relevant examples and mathematics at more appropriate levels. Perhaps the statistical journals could use on-line links to present multiple subject fields and associated computer code (in several computer languages) to illustrate their new techniques.

   Worse are those examples of known techniques re-invented, wasting person-hours and financial resources.

   One problem is the relative lack of statistical journals indexed in Medline as opposed, to say, Science Citation, leading medical researchers to overlook relevant publications in statistical journals.

   Helpful would be subject-area experts more capable of readily understanding the statistical techniques.

6. My team works in energy grid maintenance, and we're the first group to do this on a large scale, in NYC. We know it's an important area given the state of our grid. We use statistical technology developed originally for problems in web search and information retrieval. There is a lot of funding going to theoretical work and simulations on smart meters, but we are actually fixing the power grid, and there are no funding opportunities that we've found for energy grid maintenance. Luckily we are funded to some extent by the power company in NYC since we are actually helping them.

Currently there is work done at UCLA and in other places where they are using earthquake prediction models to predict crime. We noticed that the same methods are used in neuroscience, and now we're using them for energy grid maintenance and medical outcome prediction. It is essentially the same statistical model used in 5 totally different areas.
It would be nice if NSF would encourage cross-disciplinary work, and new application areas.

7. In synthetic biology computer scientists, engineers, users and developers of microarrays are working together. Expertise in statistics is needed in dealing with large data matrices.
    - Genomic analyses, GWAS (Genome-wide Association Studies) also need statistical inputs.
    - Economics seems to lack adequate use of meta-analyses. As a result, analyses appear de novo which should incorporate the best of prior empirical evidence.
    - Social science generally and ecology generally should incorporate methods from biomedical research.

8. Generation of massive data sets is now becoming a norm rather than exception in almost all areas of scientific activity particularly due to high-throughput technologies. For instance genomic research has been generating data, through sequencing of different species of crops, livestock, fishery etc. as well as of humans, of the order of terabytes or even petabyte. The computational as well as statistical challenges to maintain, process, and interpret them and to integrate them with other similar sets are enormous.

    To quote one example from human disease and health, asthma datasets were collected under Severe Asthma Research Program (SARP) involving 543 asthma patients with genotype data for 34 SNPs within or near the IL-4R gene that spans a 40-kb region on chromosome 16. The phenotypic data included 53 clinical traits related to severe asthma such as age of onset, family history, and severity of various symptoms. The objective was to determine whether any of the SNPs in the region were associated with a subnetwork of correlated traits rather than an individual trait. The structure of data requires the use of sparse regression methods such as ridge regression, lasso, elastic net etc. considering only a single trait. For the given example on asthma with 53 clinical traits, the data were first expressed as a quantitative trait network (QTN) along with subnetworks of quality of life, asthma symptoms, and lung physiology with a view to find out the common SNPs associated with the disease. Kim and Xing (2009) extended the lasso technique to include a fusion penalty first described by Tibshirani et al. (2005) in classical regression problems with time element. The results led to a new hypothesis that two SNPs in this set as well as a known SNP might be jointly associated with the same subset of traits for lung physiology.

    The above example illustrates how statistical methodology developed elsewhere can be fruitfully adopted to a new situation resulting in new findings which may be obscure to a researcher in his own field. Several instances of this type can be quoted to support the usefulness of statistical science in a disciplinary field.
    References

    - Kim, S. and Xing, E. P. (2009). Statistical estimation of correlated genome associations to a quantitative trait network. PLoS Genet. 5(8): e 1000587. doi: 10.1371/joural.pgen. 1000587.
    - Tibshirani, R., Saunders, M., Rosset, S. ,Zhu, J., and Knight, K. (2005). Sparsity and Smoothness via the fused lasso. Journal of Royal Statistical Society, Series B 67: 91-108.

9. There are many examples of statistical science methodologies that are under-utilized in ecology, evolution and organismal biology. Some include:
    a. Improved methods for model selection, including methods beyond Akaike Information Criterion and methods for incorporating uncertainty due to model selection into final results and forecasts.
    b. Advances in computational methods such as improved Markov chain Monte Carlo algorithms, Approximate Bayesian Computation (ABC), Integrated Nested Laplace Approximation (INLA), and others.
    c. Improved synthesis and application of Generalized Linear Mixed Models and Generalized Additive Mixed Models.
    d. Application of methods such as Causal Inference and Generalized Estimating Equations that have not been used much in ecology.
    e. Methods for testing hypotheses with "big data", including control of error and/or false discovery rates.

Responses to Question 5: What are some examples of simultaneous development of statistical science methods for different fields that might benefit from cross-fertilization?

1. Operations Research/Management Sciences, Economics/Econometrics, Business Analytics, Machine Learning/Data Mining, Big Data, and Data Sciences.

2. String theory being used for quantum effects, CERN particle statistics which could be useful tools for biological or chemical sciences.

3. The research areas in statistics that I think need more attention and resources include (1) exploratory data analysis, visualization and modeling, (2) statistical learning, e.g., statistical methods for data mining, (2) high-dimensional data analysis, especially when sample size is smaller than variable dimension.

4. Same as response for question 5: The interdisciplinary work involving the intersection of geography, urban planning, public health and behavioral economics have the most promise for us.

5. The big-data techniques, once their statistical capabilities are known (robustness, etc.), need to be quickly pushed to applicable areas before the techniques are re-invented by those applicable areas.

   Work at CERN in spotting rare events among very large data sets may be a candidate for more wide-spread applications.

6. Social network analysis, network sampling applied in public health surveys. Bayesian methods for imputation in network sampling of human populations. Times series methods for forecasting morbidity, mortality and live-birth counts.

7. Mutual sharing of elements common to Geographic Information Science (GIS) and Statistical Science; e.g., emphasis upon the individual rather than groups and location (latitude/longitude), layering of data and locational analytics.

   Quantum computing and exploratory data analysis methods explicitly aimed at mega databases are applicable across vast application ranges.

8. Ideas that were developed for recommendation systems also find applications in image processing (inpainting), due to the connection to matrix completion. Again, this is just one example. The EM (expectation maximization) algorithm continues to find new applications and can even influence researcher's thinking about his or her problem area. And all this is only possible due to the continuing progress in computing methods and computing power.

9. Some examples are:
   1. Computational methods for maximum likelihood estimation of hierarchical (mixed) models. For example, Lele et al. (2007) and Jacquier et al. (2007) simultaneously introduced the same method in ecology and in economics, respectively.
   2. Theory and methods for generalized linear mixed models, including related generalized estimating equations.
   3. Theory and methods for nonlinear dynamic systems.
   4. Theory and methods for spatial random field models.
   5. Modeling with zero-inflated data distributions or compounded distributions.

10. Economists and statisticians tend to approach similar problems – both have methods for causal inference and both also deal with time series data.

Responses to Question 6: Are there research areas in statistical science that, with sufficient funding support, could spur significant advances in science? If so, please specify.

1. Clearly an answer to (6) is methods for analyzing large and/or highly-structured data sets that arise in application.

2. Yes, I do think the area of "big data" will be beneficial to us in social work and will cut across disciplines well especially if it integrates methods that better understands person-in-environment.

3. In response to your question number 6, I suggest that statistical research supporting the broad field of Computational Biology including Genomics, Proteomics, and Bio Informatics has the potential to spur significant scientific advances to improve global health and well-being. These advances can directly influence the development of more effective disease treatments, drugs, and seed varieties.

   Computational Biology, “the development and application of data-analytical and theoretical methods, mathematical and statistical models, and computational simulation techniques to the study of biological systems,” involves the collection of very large amounts of data from different

fields and the application of computationally intensive methods to mine the data to understand the relationships between biological and clinical outcomes.

This involves the development and implementation of:

- Data Mining, Visualization, and Pattern Recognition Algorithms, Statistical Models and Methods, and Simulation Modeling and Analysis to assess the relationships among the elements of very large data sets,
- Predictive models integrating top-down hypothesis driven models with the bottom-up data-driven models,
- Tools and databases to store, manage, and access very large and disparate data sets.

Statistical Science and Statisticians are well-positioned to make significant contributions to this collection of related and rapidly developing new fields focused on recent advances in Genomics and Information Technology.

4. I think the area of statistical learning (classification, data mining, prediction) offers huge opportunities for statisticians to make substantive contributions to spur significant advances in science. One popular application area is "personalized medicine," but there are many others. I believe that DoD (DTRA) sponsors such research in its counter-terrorism efforts, and I suspect there are top statisticians who tap that resource. NSF could sponsor such research, too, if not already doing so.

5. I think so, especially in experimental design…if we could have ways to get to an answer earlier this would be very helpful. More research on using priors along with new data to show inter-relationships would be very helpful.

6. The non-independence of big-data observations may be under appreciated.

7. Statistical methods for software robotics. I believe my responses to questions 4 and 5 indicate other areas where additional funding support could spur significant advances in science.

8. Perhaps we should encourage areas in statistical science that would allow a better *understanding* of data generally. For instance, creating statistical models that are interpretable, meaning models that allow humans experts insight into their data, in order to better be able to solve scientific problems. We have found that if predictive models are optimized very heavily to be interpretable, one can really gain a lot of insight into the problem and data.

9. Longitudinal analysis, with attention to intervention effects, e.g., adjustment to time-series for changes in frame and data collection procedures.

   Use of a blend of analysis and simulation-oriented techniques in (a) determining statistical parameters and in conveying their meaning, (b) estimating reliability implications, (c) stagewise quality control.

10. Most of the advances foreseen would result from greater phenomenological understanding, and increased predictive ability, although the application of statistical techniques to appropriately designed knowledge-based systems may also provide quantum benefits.

11. Improved support and training for statistical research would be a boon to fields such as ecology, evolution and organismal biology. Some examples of statistical research needs from these application domains include:
    a. In phylogenetics, the ability to collect genetic data is no longer a bottleneck in the scientific process, and instead the current limiting factor is the ability to make robust inference from the large datasets becoming available.
    b. Interpretation of data from high throughput proteomic, phenomic, and other "–omics" is limited by the availability of robust methods for analysis.
    c. Methods for high-dimensional categorical and compositional data that arise commonly in biology.
    d. New methods for analyzing large and/or meta-analytic datasets of the types listed above.
    e. Cross-disciplinary work on the role of statistical science in application domains that interact with social sciences and management.

12. Validation of data mining