Statistics

This chapter contains the report of the Subpanel on Statistics of the CUPM Panel on a General Mathematical Sciences Program, reprinted with minor changes from Chapter VI of the 1981 CUPM report entitled RECOMMENDATIONS FOR A GENERAL MATHEMATICAL SCIENCES PROGRAM.

Introductory Course

Statistics is the methodological field of science that deals with collecting data, organizing and summarizing data, and drawing conclusions from data. Although statistics makes essential use of mathematical tools, especially probability theory, it is a misrepresentation of statistics to present it as essentially a subfield of mathematics.

The Statistics Subpanel believes that an introductory course in probability and statistics should concentrate on data and on skills and mathematical tools motivated by the problems of collecting and analyzing data. The traditional undergraduate course in statistical theory has little contact with statistics as it is practiced and is not a suitable introduction to the subject. Such a course gives little attention to data collection, to analysis of data by simple graphical techniques, and to checking assumptions such as normality.

The field of statistics has grown rapidly in applied areas such as robustness, exploratory data analysis, and use of computers. Some of this new knowledge should appear in a first course. It is now inexcusable to present the two-sample t-test for means and the F-test for variances as equally legitimate when a large literature demonstrates that the latter is so sensitive to nonnormality as to be of little practical value, while the former (at least for equal sample sizes) is very robust (e.g., see Pearson and Please, *Biometrika* 62 (1975), pp. 223-241, for an effective demonstration). However, the Statistics Subpanel does not believe that a course in "exploratory data analysis" is a suitable introduction to statistics, nor does it advocate replacing (say) least squares regression by a more robust procedure in a first course. But it does think that new knowledge renders a course devoted solely to the theory of classical parametric procedures out of date.

While the Statistics Subpanel prefers a two-semester introductory sequence in probability and statistics, enrollment data shows that most students take only a single course in this area. The course proposed below gives students a representative introduction to both the data-oriented nature of statistics and the mathematical concepts underlying statistics. These broad objectives raise several issues that require preliminary comment. One year of calculus is assumed for this course. The course should use Minitab or a similar interactive statistical package.

The Place of Probability

Probability is an essential tool in several areas of the mathematical sciences. It is not possible to compress a responsible introduction to probability and coverage of statistics into a single course. The Statistics Subpanel therefore recommends that probability topics be divided between the courses on probability and statistics, discrete methods, and modeling/operations research as follows:

- *Probability and statistics course:* Axioms and basic properties; random variables; univariate probability functions and density functions; moments; standard distributions; Laws of Large Numbers and Central Limit Theorem.
- *Discrete methods course:* Combinatorial enumeration problems in discrete probability.
- *Modeling/operations research course:* Conditional probability and several-stage models; stochastic processes.

This division is natural in the sense that the respective parts of probability are motivated by and applied to the primary concerns of these courses.

Alternative Arrangements

The subpanel is convinced that two semesters are required for a firm introduction to both probability and statistics. Many institutions now offer such a two-semester sequence in which probability is followed by statistics. The subpanel prefers this structure. In this sequence the statistics course should be revised to incorporate the topics and flavor of the data analysis section of the proposed unified course. With probability first, added material in statistics can also be covered, such as Neyman-Pearson theory, distribution-free tests, robust procedures, and linear models.

Institutions will vary considerably in their choice of material for this statistics course, but the subpanel reiterates its conviction that the traditional "theory-only"

statistics course is not a wise choice. If experience shows that many students drop out in the middle of a two-course sequence, the unified course outlined below should be adopted, followed by one of the elective courses suggested in Section 3 of this chapter.

Instructor Preparation

Since the recommended outline is motivated by data and shaped by the modern practice of statistics, many mathematically trained instructors will be less prepared to teach this course than a traditional statistical theory course. Growing interest in "applied" statistics has, of course, led many instructors to broaden their knowledge. Some background reading is provided for others who wish to do so. The publications listed here contain material that can be incorporated in the recommended course, but none is suitable as a course text. In order of ascending level:

1. Tanur, Judith, *et al.*, eds., *Statistics: A Guide to the Unknown,* Second Edition, Holden-Day, 1978.
   An elementary volume describing important applications of statistics and probability in many fields of endeavor.
2. Moore, David, *Statistics: Concepts and Controversies,* W.H. Freeman, 1979.
   A paperback with good material on data collection, statistical common sense, appealing examples, and the logic of inference.
3. Freedman, David; Pisani, Robert; Purves, Roger, *Statistics,* W.W. Norton, 1978.
   A careful introduction to elementary statistics written with conceptual richness, attention to the real world, and awareness of the treachery of data.
4. Mosteller, Frederick and Tukey, John, *Data Analysis and Regression,* Addison-Wesley, 1977.
   Good ideas on exploratory data analysis, robustness and regression.
5. Box, George; Hunter, William; Hunter, J. Stuart, *Statistics for Experimenters,* Wiley, 1978.
   Applied statistics explained by experienced practical statisticians. Some specialized material, but much of the book will repay careful reading.
6. Efron, Bradley, "Computers and the Theory of Statistics: Thinking the Unthinkable," *SIAM Review,* October, 1979.
   A superb article on some new directions in statistics, written for mathematicians who are not statisticians.

Course Outline

I. Data (about 2 weeks)

- *Random sampling.* Using a table of random digits; simple random samples, experience with sampling variability of sample proportions and means; stratified samples as a means of reducing variability.
- *Experimental design.* Why experiment; motivation for statistical design when field conditions for living subjects are present; the basic ideas of control and randomization (matching, blocking) to reduce variability.

COMMENTS: Data collection is an important part of statistics. It meets practical needs (see Moore) and justifies the assumptions made in analyzing data (see Box, Hunter and Hunter). Experience with variability helps motivate probability and the difficult idea of a sampling distribution. Students should see for themselves the results of repeated random sampling from the same population and the variability of data in simple experiments such as comparing 3-minute performance of egg timers (see W.G. Hunter, *American Statistician,* 1977, pp. 12-17, for suggestions).

II. Organizing and Describing Data (about 2 weeks)

- *Tables and graphs.* Frequency tables and histograms; bivariate frequency tables and the misleading effects of too much aggregation; standard line and bar graphs and their abuses; box plots; spotting outliers in data.
- *Univariate descriptive statistics.* Mean, median and percentiles; variance and standard deviation; a few more robust statistics such as the trimmed mean.
- *Bivariate descriptive statistics.* Correlation; fitting lines by least squares. If computer resources permit, least-square fitting need not be restricted to lines.

COMMENTS: In addition to simple skills, students must be trained to look at data and be aware of pitfalls. Freedman, Pisani and Purves have much good material on this subject, such as the perils of aggregation (pp. 12-15). The impressive effect on a correlation of keypunching 7.314 as 731.4 should be pointed out. Simple plots are a powerful tool and should be stressed throughout the course as part of good practice.

III. Probability (about 4 weeks)

- *General probability.* Motivation; axioms and basic rules, independence.
- *Random variables.* Univariate density and probability functions; moments; Law of Large Numbers.

- *Standard distributions.* Binomial, Poisson, exponential, normal; Central Limit Theorem (without proof).
- *More experience with randomness.* Use in computer simulation to illustrate Law of Large Numbers and Central Limit Theorem.

COMMENTS: Probability must unavoidably be pressed in a unified course that includes data analysis. Instructors should repeatedly ask "What probability do I need for basic statistics?" and "What can the students learn within about four weeks?" It is certainly the case that combinatorics, moment generating functions, and continuous joint distributions must be omitted. Some instructors may be able to cover conditional probability and Bayes' theorem in addition to the outline material.

IV. Statistical Inference (about 6 weeks)

- *Statistics vs. probability.* The idea of a sampling distribution; properties of a random sample, e.g., it is normal for normal populations; the Central Limit Theorem.
- *Tests of significance.* Reasoning involved in alpha-level testing and use of P-values to assess evidence against a null hypothesis; cover one- and two-sample normal theory tests and (optional) chi-square tests for categorical data. Comment on robustness, checking assumptions, and the role of design (Part I) in justifying assumptions.
- *Point estimation methods.* Method of moments; maximum likelihood; least squares; unbiasedness and consistency.
- *Confidence intervals.* Importance of error estimate with point estimator; measure of size of effect in a test of significance.
- *Inference in simple linear regression.*

COMMENTS: A firm grasp of statistical reasoning is more important than coverage of a few additional specific procedures. For much useful material on statistical reasoning such as use of the "empirical rule" to assess normality, see Box, Hunter and Hunter. *Don't* just say, "We assume the sample consists of iid normal random variables." Applied statisticians favor P-values over fixed alpha tests; a comparative discussion of this issue appears in Moore.

RECOMMENDED TEXTS

The Subpanel is not aware of a text at the post-calculus level that fits the recommended outline closely. Instructors should seriously consider adopting a good post-calculus statistical methods text rather than a theoretical statistics text. A methods text is more likely to have examples and problems which have the ring of truth. Moreover, most instructors will find it easy to supplement a methods text with mathematical material and problems familiar from previous teaching. It is much harder to supply motivation and realistic problems, and it is psychologically difficult for both the teacher and student to skip much of the probability in a mathematical statistics text.

The following books are possible texts or reference material for the course described above. All of these have essentially the same shortcoming of being too "unmathematical." The appropriate combination of level of sophistication and content is not now available under a single cover. The class of books below fall in the "intermediate" level between an elementary statistics course and a first course in mathematical statistics.

1. Box, George; Hunter, William; Hunter, J. Stuart, *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building,* John Wiley & Sons, New York, 1978.
2. Moore, David and McCabe, George, *Introduction to the Practice of Statistics,* Freeman, San Francisco, 1989.
3. Ott, Lyman, *An Introduction to Statistical Methods and Data Analysis, Second Edition,* Duxbury, Boston, 1984.
4. Neter, John; Wasserman, William; Whitmore, G.A., *Applied Statistics,* Allyn and Bacon, Boston, 1978.

Additional Courses

Probability and Statistical Theory

CONTENT: Distribution functions; moment and probability generating functions; joint, marginal and conditional distributions; correlations; distributions of functions of random variables; Chebyschev's inequality; convergence in probability; limiting distributions; power test and likelihood ratio tests; introduction to Bayesian and nonparametric statistics; additional regression topics.

COMMENT: This course is designed to complete the traditional probability-then-statistics sequence. Since the students have already completed a semester of study, they should be capable of tackling a good text on mathematical statistics such as the one by DeGroot or by Hogg and Craig. The book by Bickel and Doksum is a little more difficult than the other two, and the teacher would have to supplement it with the topics in probability.

TEXTS:

1. Mendenhall, William; Schaeffer, Richard; Wackerly, Dennis, *Mathematical Statistics with Applications, Second Edition,* Duxbury, Boston, 1981.
2. Larsen, Richard and Marx, Morris, *An Introduction to Probability and its Applications,* Prentice-Hall, Englewood Cliffs, N. Jers., 1985.
3. DeGroot, Morris M., *Probability and Statistics,* Addison-Wesley, Reading, Mass., 1975.
4. Hogg, Robert and Craig, Allen, *Introduction to Mathematical Statistics,* Macmillan, New York, 1978.

Applied Statistics

CONTENT: This course uses statistical packages to analyze data sets. Topics include linear and multiple regression; nonlinear regression; analysis of variance; random, fixed and mixed models; expected mean squares; pooling, modifications under relaxed assumptions; multiple comparisons; variance of estimators; analysis of covariance.

COMMENT: The new introductory course will probably attract more students from other fields than the traditional probability-then-statistics course. This course is an excellent follow-up for such non-mathematical sciences students. Its topics are among the more widely used statistical tools. Students should be expected to use a statistical computing package such as Minitab of SPSS for many of the analyses. The book by Miller and Wichern is a possible text for this course.

TEXTS:

1. Miller, Robert and Wichern, Dean, *Intermediate Business Statistics,* Holt, Rinehart and Winston, New York, 1977.
2. Neter, John; Wasserman, William; Kutner, Michael, *Applied Linear Statistical Models, Second Edition,* Irwin, 1985.
3. Morrison, Donald, *Applied Linear Statistical Methods,* Prentice-Hall, Englewood Cliffs, N. Jers., 1983.

Probability and Stochastic Processes

CONTENT: Combinatorics; conditional probability and independence; Bayes theorem; joint, marginal and conditional distributions; distribution functions; distributions of functions of random variables; probability and moment generating functions; Chebyschev's inequality; convergence in probability; convergence in distribution; random walks; Markov chains; introduction to continuous-time stochastic processes.

COMMENT: This is a fairly standard course and a number of texts are available. The book by Feller is a classic but covers only discrete probability. The book by Olkin, Gleser and Derman is at a slightly lower level and is more "applied" but will require the instructor to provide some supplementary materials. The book by Chung is excellent but must be read with a "grain of salt." The book by Breiman is also excellent but expects much of its reader. A new book by Johnson and Kotz also looks interesting but is restricted to discrete probability. The books by Chung, Feller and Breiman are difficult for the average student.

TEXTS:

1. Olkin, Ingram; Gleser, Leon J.; Derman, Cyrus, *Probability Models and Applications,* Macmillan, New York, 1980.
2. Larsen, Richard and Marx, Morris, *An Introduction to Probability and its Applications,* Prentice-Hall, Englewood Cliffs, N.J., 1985.
3. Ross, Sheldon, *A First Course in Probability, Second Edition,* Macmillan, New York, 1984.
4. Chung, Kai Lai, *Elementary Probability Theory with Stochastic Processes,* Springer-Verlag, New York, 1974.
5. Feller, William, *An Introduction to Probability Theory and Its Applications, Volume I,* John Wiley & Sons, New York, 1950.

Preparation for Graduate Study

There are a large number of career opportunities for statisticians in industry, government and teaching. For example as of 1977, the Federal Government employed over 3500 statisticians, plus 3500 statistical assistants and numerous other employees performing statistical duties but classified in different job series. A recent report by the U.S. Labor Department, reprinted in the *New York Times* National Recruitment Survey, predicts an increase of 35% in the demand for statisticians during the 1980's. This compares to a predicted increase of 9% for mathematicians and 30% for computer specialists.

Preparation for a career in statistics usually involves graduate study. An undergraduate major in statistics, computer science, or mathematical sciences is the recommended preparation for graduate study in statistics. It is desirable for such a major to include solid courses in matrix theory and real analysis, in addition to courses in probability and statistics. Most statistics graduate programs require matrix algebra and real analysis for fully matriculated admission. Either one of the sample programs in the report of the General Mathematical Sciences Panel in the first chapter would be adequate preparation for graduate study in statistics. However,

major A is preferable to major B, and both should include at least one follow-on elective in probability and statistics.

In addition to courses in the mathematical sciences, a student preparing for graduate study in statistics should:

- Study in depth some subject where statistics is an important tool (physics, chemistry, economics, psychology, ...). In fact, a double major should be considered.
- Take as many courses as possible which are designed to enhance his or her communication skills. Statisticians in industry and government are often called upon to provide written reports and critiques; consulting requires clear oral communications.

A detailed discussion of preparation for a statistical career in industry can be found in [1] A similar report, [2], discusses preparation for a career in government.

1. Preparing Statisticians for Careers in Industry: Report of the ASA Section on Statistical Education. *The American Statistician,* 1980, pp. 65-80.
2. Preparing Statisticians for Careers in Government: Report of the ASA Section on Statistics in Government. Paper presented at the American Statistical Association meeting in August, 1980.

Panel Members:

RICHARD ALO, CHAIR, Lamar University.
RICHARD KLEBER, St. Olaf College.
DAVID MOORE, Purdue University.
MIKE PERRY, Appalachian State University.
TIM ROBERTSON, University of Iowa.