A GRADUATE COURSE IN BASIC STATISTICAL ANALYSIS FOR MAJORS OR MINORS IN STATISTICS

BY

R. L. Anderson,
Institute of Statistics,
North Carolina State College

Institute of Statistics
Mimeo Series No. 111
July 7, 1954

1. INTRODUCTION: The Institute of Statistics of the Consolidated University of North Carolina offers advanced degrees in Experimental Statistics at North Carolina State College and in Mathematical Statistics at the University at Chapel Hill. In order to more adequately coordinate our teaching program, it was decided to require our beginning graduate students to take a uniform set of basic courses in statistical theory and analysis. Only those graduate students who had secured similar work before coming to North Carolina were to be excused from taking one or both of these courses. If possible, it was hoped that these courses would also serve as the basic statistical training for graduate students minoring in statistics. This latter point was of particular importance for the Experimental Statistics department, because we have been encouraging the training of statistical consultants in many fields. However, the major emphasis in developing the two basic courses was on the content needed for majors in statistics.

A committee from the staff of the Institute drew up a proposed outline for the two basic courses. The basic theory course was taught by R. C. Bose and the basic analysis course by this writer on both campuses for the academic year 1953-54. In Table 1 are listed the books which were available for reference purposes in these basic courses. We drew up another list of about sixty other books, which were recommended to be in the libraries for occasional use by the students. Lecture notes for the theory course were prepared by Dr. Bose and for the analysis course by R. J. Hader and this writer.

An outline of the basic theory course is presented in Table 2. Dr. Bose's lectures were superb and showed that a tremendous amount of time and effort had been devoted to their preparation. Almost all students who were able to take both courses did very well in both of them. On the whole, it appeared that the original objectives of the committee had been achieved, namely to develop two integrated sequences for majors in statistics. There was insuf-

ficient time to cover the last three topics listed in Table 2. It may be that more than four lecture hours per week are needed to present all the material. However, it is possible that on the basis of this one year's experience, the course will be reorganized so that all topics can be covered in the allotted time.

The remainder of this paper will be devoted to a discussion of the Basic Analysis course, an outline of which is presented in Table 3. Since this was a course in analysis, we decided to allocate four hours a week to supervised lab work which would enable the students to study sampling distributions by drawing random samples from known populations and to analyze actual data. The laboratory sessions were to be preceded by two lecture hours each week. Unfortunately, two hours a week did not provide sufficient time to present in lectures the materials and explanations needed for the laboratory work. I had planned to use some of the laboratory time to catch up on lectures but found that the sampling experiments required all the regular lab time; hence, some topics were omitted from this course and others only mentioned briefly. On the basis of this experience, it appears that three lecture hours per week are needed.

No regular textbook was used for the descriptive statistical tools. Dixon and Massey was used for the sampling experiments and general methods of estimation and tests of significance, including non-parametric methods and sequential tests. I lectured from Cochran's sampling book for that topic and used my part of Anderson and Bancroft for regression analysis, analysis of variance and components of variance.

It was decided that considerable time should be devoted to discussions of descriptive statistics, because of the large number of government and business personnel who collect and analyze census, social and business data. The students were to be introduced to the problem of making inferences from samples by drawing random samples from known populations. Then the course was to be completed by analyzing actual survey and experimental data and learning to efficiently design

surveys and experiments.

2. DESCRIPTIVE METHODS: Now let me go over the course in the order presented in Table 3. We first prepared a preliminary mathematics quiz on logarithms, summations and solution of equations; this was used to indicate if some of those students who were not taking the Basic Theory course might be weak enough in mathematics to require some review while taking the analysis course. The laboratory sessions for the first week plus the regular lecture period were spent on the introductory material, including a brief history of statistical development and a discussion of present uses of statistical analysis and methods and sources of collecting data. In the future it seems that considerable time would be saved by:

(1) mimeographing much of this discussion and distributing it to the students;

(2) assigning outside reading in the cited references.

Next methods of describing populations were presented, first for attribute or classification data. Most students are already familiar with cross tabulations of census and market data; hence, this seemed a logical starting point. Various terms needed to be defined, such as frequency tables, relative frequencies, frequency distributions, marginal totals and distributions, conditional distributions, dichotomy, and multichotomy. The notation of Yule and Kendall was used for much of this material. Various measures of associations were considered. Many measures of association are available, but there does not seem to be general acceptance of any single measure. This writer feels that, in an enthusiasm to develop methods of making inferences about hypothetical populations on the basis of small samples or experiments, we have neglected to improve on methods of describing finite population data. Too little attention has been paid to the problem of connecting descriptive measures with the purposes of the investigation.

Next we considered the more definite problem of describing variable (or scaled) data, first that of finite populations. These data are essentially of

two types:

discrete, e.g. size of family and number of rooms in a house;

continuous, e.g. farm acreage and income.

If the number of individuals in the population is large, a frequency table is useful to indicate if there are any concentrations of individuals. But if the data are continuous, one has the problem of determining class intervals and the number of classes. Most advice on these matters is rather vague; I usually advise that one use from 8-20 classes with rather simple class intervals. One wants to smooth out small irregularities. It is important to emphasize that the midvalue of the class interval should represent the class; this was demonstrated with some acreage data for midwestern farms, which tend to concentrate at multiples of 80 acres.

The usual measures of location and dispersion were then discussed, including index numbers. One needs to indicate when certain measures seem better than others. There is too much of the attitude that the arithmetic mean is the proper measure of location. This is generally true when the results are used to obtain population totals, but I feel that there are many circumstances in which the mean is not a very good indicator of location. The same holds for a discussion of measures of dispersion. The standard deviation has some real meaning for near-normal data, but not for many decidedly non-normal populations.

After attempting to discuss those distributions which are typified by a mathematical density function, I have concluded it is best to treat them as hypothetical distributions which apply more or less closely to many actual sets of data. One can explain that each of these has a certain logical theoretical basis for being considered in given experimental situations. It is in connection with this aspect of the course that a parallel course of theory is quite useful; hence, most of the development of the theory of these distributions can be omitted in the analysis course in order to concentrate on examples of data which closely approx-

imate the theoretical distributions.

It is important to emphasize that these special distributions are characterized by certain parameters. Sometimes values of these parameters can be obtained by a theoretical study or on the basis of a priori reasoning; however, in general it is necessary to secure estimates by conducting an experiment or survey. Here is an opportunity to connect description with inference. The student soon realizes that he cannot study an entire infinite population; and that sampling is necessary for many large finite populations. In this connection, the teacher can point out the existence of tables and mathematical formulas which are quite useful in summarizing data if the populations closely approximate one of these special distributions. This can be followed by a discussion of transformations, and the Central Limit Theorem.

When one comes to the problem of describing the relationship between two continuous variates, he can present a two-way frequency table, but the problem of a single measure of association is solved only for the bivariate normal. Again it becomes useful to study the problem from a theoretical point of view; certainly the student should be aware of the difference between independence and non-correlation for non-normal variates. And if the variates are not normal, why should one use this particular cross-moment as the measure of association? I was forced to state that these results applied to near-normal data, with no clear statement of the meaning of "near".

Conditional distributions, which are important in regression analysis, can be demonstrated with a two-way frequency table. One usually has trouble here because of the small frequencies on the tails; hence, it may be best to use unequal class intervals in order to obtain about the same frequency for each class. Deviations from linearity reflect either the finiteness of the data or its essential non-normality. Bivariate normal models are often useful in studying the basic ideas of conditional distributions.

I doubt if regression analysis with one or more variables fixed should be introduced at this stage. Actually this is simply an extension of normal theory. Multivariate analysis should be mentioned, but I would not stress it in an introductory course.

3. SAMPLING EXPERIMENTS AND STATISTICAL INFERENCE: The concepts of statistical inference were introduced by having each member of the class draw samples from known populations. The populations studied were:

(1) The random normal numbers and the random two-digit numbers in Dixon and Massey. The latter were used to represent both a binomial and a uniform population.

(2) A set of 160 acreages of midwestern farms to represent a finite multimodal population.

Each student was asked to draw 3 samples of 10 from each of these populations. I made such a draw first and had the results distributed to the class as an example of the required computing. My samples from the random numbers are presented in Table 4. A summary of some of the results of the class sampling from this simulated uniform population are presented in Table 5. In order to study the effect of increased sample size, the three samples of 10 were composited to form a sample of 30.

A binomal population with $p = 0.3$ was created by making the digits 0, 1 or 2 successes and the other seven digits failures. Hence, a sample of 20 was obtained in each set of 10 two-digit numbers drawn from the random numbers. The first digits from each number were used to form two samples of size 10 from each student. The frequency table for samples of size 10 was:

X	0	1	2	3	4	5	6	7	Sum	Mean	Var.
f	5	21	35	37	26	12	1	1	138	2.746	1.855

Previously the acreage data had been used to demonstrate sampling theory for means to a class of agricultural economists; 200 samples of 10 each were obtained.

It is quite evident that only tentative conclusions can be derived on the basis of the limited number of samples obtained in the laboratory classes. In order to demonstrate the results for a large number of samples, we used a set of IBM cards punched with the Mahalanobis random normal numbers.* Certain obvious errors in these tables had been deleted; so, that we had a total of 10,390 cards. Samples of 10 and 30 were obtained, giving 1,039 samples of 10 and 346 samples of 30. These samples were used to compute the same statistics as in class. Frequency distributions were derived for the various estimates of location and dispersion and the t and χ^2 testing statistics.

The students were given all the composited data: the IBM normal data, the previous study for the acreage data, and the class results for the uniform and binorial data. Each student was asked to write a brief term paper on the uses of empirical sampling and the results of the above sampling. On the basis of this experience with using empirical sampling to demonstrate various concepts of statistical inference, these comments seem pertinent:

(1) Many students tend to get lost in the computing and "fail to see the forest for the trees". In the future, I would spend more time briefing them and on running small sampling projects from small populations before starting the mass computations. Admittedly one of the major difficulties was in compositing the results of sampling done at two places.

(2) Contrary to the advice I have received from others, most students seemed to grasp the results of sampling done outside of class. I feel that we would have accomplished much more if our IBM calculations had been available at the start.

(3) The greatest benefit seemed to be in a better appreciation of the meaning of confidence intervals and Type I and II errors. We demonstrated the latter by using the normal data with non-zero means.

(4) The purposes of empirical sampling were best understood by those students who also were taking the theory course.

---

*P. C. Mahalanobis, S. S. Bose, P. R. Roy and S. K. Banerji, "Tables of Random Samples from a Normal Population", Sankhyā 1:289-328 (1934).

After the students had completed the sampling studies, they were assigned enough applied problems to acquaint them with the actual uses of the estimating and testing procedures.

A sampling experiment was also set up to study the use of χ^2 in tests of significance with enumeration data. The description of this experiment is given in Table 6. The results of 194 samples indicated that the correction for continuity for these particular data severely over-corrected, i.e. the 5% point became the 2% point, and no samples were significant at the 1% point. As expected, the uncorrected χ^2 for one degree of freedom gave a few too many significant results, i.e. the 5% point was the 8% point and the 1% point was the 2% point. The 10% point gave us 13% significant results for the uncorrected χ^2 and 6% for corrected χ^2. The results for χ^2 with three degrees of freedom were quite satisfactory. Finally, on the question of power, a reversal of the order of magnitude in Table 6 would not be likely to be detected; about 10% were significant when using the 5% uncorrected χ^2 critical region and 5% significant in the 1% critical region.

4. THE ANALYSIS OF DATA AND DESIGN OF EXPERIMENTS AND SURVEYS: The students were asked to read, and I discussed, the χ^2 article by Cochran in the September, 1952 Annals of Mathematical Statistics. A set of exercises on the analysis of enumeration data was prepared for subsequent laboratory sessions. The following points were stressed in the lectures:

(1) Assumptions in the use of χ^2

(2) Construction of single degree-of-freedom comparisons

(3) Effect of estimating parameters in goodness-of-fit

(4) Exact tests in 2 x 2 tables

The lectures on control charts, non-parametric (distribution-free) methods, and sequential testing were patterned after the presentation in Dixon and Massey. Exercises were assigned from this book. The Wilcoxon rank-sum test was added to the non-parametric tests presented in Dixon and Massey. No sampling experi-

ments were conducted in the laboratories, because it became apparent that there would not be sufficient time to analyze actual data and conduct sampling experiments for the remainder of the course. I felt that the principles of sampling had been rather thoroughly gone over by this time. It was deemed advisable to restrict subsequent laboratory periods to the analysis of actual experimental survey data and to principles of designing experiments and surveys.

In the lectures on sequential testing, these points were emphasized:

(1) A rationale was presented for the limiting ratios in the testing procedure, i.e. one accepts H_0 whenever the ratio of the probability of the sample given H_0 is true (p_0) to that given H_1 is true (p_1) is at least as great as the ratio of the predetermined probability of accepting H_0 given H_0 is true ($1 - \alpha$) to that given H_1 is true (β).

(2) Use as variates the number of samples, $\underline{m}$, and the number of defectives, d_m.

As indicated earlier, most of the material on surveys was taken from Cochran's book on Sampling Techniques. Exercises were taken from this book for the laboratory sessions. These included a certain number of sampling experiments from small populations, in which all possible samples were drawn. I am convinced that this is the best method of conveying the real meaning of expected values. It also shows why sampling with and without replacement are different. One lecture was devoted to definitions of terms, reasons for sampling, principles of questionaire design, principles of selecting the sample, and methods of collecting the data. Only simple random and stratified random sampling were discussed--both using linear unbiased estimates. More complicated sampling plans and estimation devices were left for courses in Survey Theory.

It should be mentioned that originally we had planned to teach the analyses of surveys and experiments at the same time. However, I decided that it is desirable to emphasize certain essential distinctions between these two types of data collection:

(1) Finite vs. infinite populations. The more I teach this subject,

the more convinced I become that there is a logical difference between sampling from a finite universe and estimating the parameters in a regression model with some hypothetical normal error term.

(2) Uncontrolled vs. controlled sources of variation. The experimenter attempts to define the scope of his project by making certain restrictions on the variability to be encountered; the survey man takes what he gets, except for some obvious stratification devices at the start. Often the survey statistician is able to impose some post-enumerative stratification.

There are other points which overlap the two fields but which are predominantly in one or the other:

(3) The survey statistician is more interested in over-all population means and totals than is the experimenter.

(4) The experimenter has been interested mainly in relationships. The survey man is beginning to face this problem but finds it much more complicated than in experiments because of the inherent disproportionality of cell frequencies.

(5) For most surveys, large sample theory is adequate.

Some points stressed in the lectures were:

(a) Notation: finite population problems.

(b) Confidence limits and estimated sample size to attain desired accuracy.

(c) Methods of allocation to strata -widespread use of proportional allocation, because most surveys are multiple-item surveys; how to determine optimum allocation.

(d) Effect of costs on sampling methods.

(e) Use of post-enumerative stratifications.

(f) Need for improved methods with analytical studies.

The remainder of the course was devoted to the analysis of variance procedures, as presented in Part II of Anderson and Bancroft. I found that even those who were not taking the theory course did quite well in this part of the course, if they were willing to spend the time required to work the assigned problems from the book. There are many reasons to explain this improvement in comprehension in this part of the course:

(1) The teacher was more familiar with the difficulties involved and helped the students where help was needed.

(2) The non-theory students were more accustomed to experimental data.

(3) The students were more interested in practical problems than in abstract developments.

In an original outline, we had allocated 16 weeks to this part of the course; actually I had only one quarter, approximately 10 weeks. Because of this curtailment of time, it was necessary to omit certain topics and the discussion of mixed models was severely curtailed.

Some introductory remarks were made on the construction of mathemcatical models to represent experimental situations. The principle of least squares was presented in as non-mathematical manner as possible. At this point an integration of theory and analysis is almost essential. I decided to explain expectation theory,, which is so important in a discussion of the analysis of variance. The essential similarity of ordinary multiple regression models and the usual analysis of variance models should be emphasized. Some of the points stressed in regression analysis were:

(1) Basic assumptions and methods of meeting them.

(2) Confidence limits for an average value and a single predicted value.

(3) Use of the Abbreviated Doolittle method of computing for multiple regression. Necessity of learning to check data throughout. Clues to computing errors.

(4) Difficulty of interpreting individual regression coefficients in multiple regression; possible instability of regression coefficients because of multicolinearity.

(5) Use of orthogonal polynomials for smoothing and interpolation; also some models are approximately polynomial in character.

(6) Determination of constants in production surface.

Next came a discussion of experimental design, with special references to Cochran and Cox, and the the March, 1947 issue of Biometrics. One needs to stress several features of designing experiments and the analysis of the results:

(1) Applicability of the results to a wider area.

(2) Importance of randomization and replication.

(3) Place of experimental design in increasing the accuracy of experimentation.

(4) Distinguish between effects and means; define the general mean, $\rightarrow$.

(5) Model construction; assumptions in the model.

(6) Relative efficiency of designs.

(7) Use of pertinent single degree of freedom comparisons, e.g. trends.

(8) Comparison of the three basic designs: completely randomized, randomized complete blocks and Latin square. Points to be considered are number of replications per treatment, size of experiment, ease of handling experiment, error degrees of freedom, plot variability, and ease of analysis if missing data or unequal plot variability.

(9) Uses of incomplete blocks designs; concept of effective replication in balanced designs as r E, where E is the efficiency factor.

(10) Distinction between field design and treatment allocation, i.e. factorials are not field designs except under confounding.

(11) Single degree of freedom approach to factorials, as well as over-all analysis.

(12) Meaning of interactions.

(13) Factorials in describing response surfaces.

(14) Analysis of disproportionate data: Abbreviated Doolittle method of computing and method of unweighted means.

Finally the course was concluded with several lectures on variance components. The points stressed were:

(1) Classification of models.

Regression or analysis of variance model

Random model { interactions; nested sampling

Mixed model { interactions; nested sampling

(2) Uses of variance component models.

(a) Population mean of prime importance and variance components mainly to put confidence limits on mean and determine most efficient sampling plan, e.g. survey sampling.

(b) Variance components of prime importance, e.g. quantitative genetics.

(c) Both mean and components of importance, e.g. mixed models for experiments to test for treatment effects.

(3) Expectations of mean squares in analysis of variance: short-cut procedures for determining coefficients.

(4) Confidence limits for variance components.

(5) Interpretation of interaction in mixed model; expectations of mean squares.

(6) Emphasis on split-plot design and stratified sampling.

In future courses, time should be allowed to discuss the following topics:

(1) Non-balanced incomplete blocks designs, especially the lattices.

(2) Missing data problems.

(3) Confounded factorials.

(4) Experiments to explore response surfaces.

(5) Covariance.

(a) Points to consider in evaluating results.

(b) Computing procedures; basic ideas.

(c) Adjusting procedures.

(d) Standard errors.

(6) Confidence limits for variance components.

(7) More on mixed models.

(8) Recovery of interblock information in incomplete blocks designs.

5. SUMMARY: This paper has outlined a course in basic statistical analysis, taught for the first time at the University of North Carolina last year. The course was designed to parallel a similar course in basic statistical theory. Most of those students who took both courses seemed to do very well; however, students taking only the analysis course without a previous course in theory were severely handicapped.

The course was divided into three parts: Descriptive methods, sampling experiments, and analysis of data and design of experiments and surveys. The first part took too long and did not interest the students. Improved descriptive tools are needed, and better methods of teaching them, at least better than were used in this first attempt.

The sampling experiments were designed to present the basic ideas of statistical inference. by drawing samples from known populations. For many of the students, the tedium of computing tended to obscure the main objectives of the study. More time needs to be spent in preliminary lectures and on complete sampling from small populations to demonstrate the meaning of expectations. We found that students were able to utilize the results of large scale sampling outside of class, e.g. IBM sampling. After some revision of the teaching procedure, I feel that the use of empirical sampling offers real promise in presenting the ideas of statistical inference. It is important to emphasize the need for this when theory is not

available.

Even though the analysis and design part of the course was curtailed, the results seemed to be highly satisfactory. It would be desirable to have some impartial observer conduct an examination on analysis and design to find out if the students learned general principles or only those procedures mentioned by the teacher.

In order to have time to present the important methods of collecting and analyzing data, the following changes are suggested:

(1) Cut down the time devoted to the introduction by distributing mimeographed materials and requiring outside reading.

(2) Condense the descriptive materials.

(3) Have results of large scale sampling experiments available before the class work begins.

(4) Introduce empirical sampling by drawing all possible samples from small populations.

(5) Be sure lectures precede the sampling, so that students have a preview of the purposes of empirical sampling.

(6) Schedule 3 lecture hours and one three hour supervised laboratory each week. Emphasize that some non-supervised laboratory work will also be expected.

One final comment: If an analysis course is designed to parallel a similar course in theory, all students should either be taking the theory course or have had it already. Otherwise it is imperative that a certain amount of theory be taught in the analysis course.

Table 1. References for Basic Statistics Courses

Anderson, R. L. and T. A. Bancroft (1952), Statistical Theory in Research, McGraw-Hill Book Co., New York.

Anderson, R. L. and E. E. Hansemen (1942), Tables of Orthogonal Polynomial Values Extended N=104, Iowa State Coll. Exp. Sta. Res. Bull. 297.

Biometrics, March 1947 (Analysis of Variance).

Biometrics, March 1951(Variance Components).

Cochran, W. G. (1953), Sampling Techniques, John Wiley and Sons, New York.

Cochran, W. G. and G. M. Cox (1950), Experimental Designs, John Wiley and Sons, New York.

Croxton, F. E. and D. G. Cowden (1939), Applied General Statistics, Prentice-Hall, Inc., New York.

Dixon, W. J. a nd F. J. Massey (1951), Introduction to Statistical Ana lysis, McGraw-Hill Book Co., New York.

Feller, W. (1950), An Introduction to Probability and Its Application, Vol. 1, John Wiley and Sons, Inc., New York.

Fisher, R. A. (1947), Design of Experiments, 4th Edition, Oliver and Boyd, Ltd., Edinburgh and London.

Fisher, R. A. (1946), Statistical Methods for Research Workers, 10th Edition, Oliver and Boyd Ltd., Edinburgh and London.

Fisher, R. A. and F. Yates (1949), Statistical Tables, Oliver & Boyd, Ltd., Edinburgh & London.

Hald, A. (1952), Statistical Tables and Formulas, John Wiley & Sons, New York.

Kendall, M. G. (1945), The Advanced Theory of Statistics, Charles Griffin & Company, Ltd., London.

Mood, A. M. (1950), Theory of Statistics, 1st Edition, McGraw-Hill Book Co., Inc. New York.

Munroe, Marshall E. (1951), Theory of Probability, McGraw-Hill Book Co., Inc. New York.

Neyman, J. (1950), First Course in Probability and Statistics, Henry Holt and Co., New York.

Smith, J. G. and A. J. Duncan (1944), Elementary Statistics and Application, McGraw-Hill Book Co., New York.

Snedecor, G. W. (1946), Statistical Methods, 4th Edition, Collegiate Press, Inc. of Iowa State College, Ames, Iowa.

Walker, H. M. (1951), Studies in the History of Statistical Methods, The Williams and Wilkins Co., Baltimore, Md.

Wilcoxon, F. (1949), Some Rapid Approximate Statistical Procedures, American Cyanamid Co.

Yule, G. U. and M. G. Kendall (1950), An Introduction to the Theory of Statistics, 14th Edition, Charles Griffin and Co., Ltd., London.

Table 2. Outline of Course in Basic Statistical Theory

Prerequisites: A previous statistics course or graduate standing

Corequisite: Advanced calculus

Credits: Five hours each semester; four lecture hours and two laboratory hours per week.

Course Summary: This course presents the theory needed in all advanced courses in statistical analysis and the fundamentals for advanced theory courses. The essential topics are: logical foundations of probability; statistical tools used to describe populations; special sampling distributions needed in theory of estimation and testing hypotheses; theory of inductive inference; linear estimation, analysis of variance and component of variance theory and problems.

Course outline

1. Permutations and combinations
2. Probability - Sets and events
   (a). Bayes' theorem
3. Random variables
   (a). Discrete and continuous
   (b). Bivariate continuous
4. Binomial distribution
5. Normal: Normal approximation to binomial
6. Poisson Distribution
7. Mean and variance
8. Introduction to basic concepts
   (a) Testing hypotheses - critical region and power
   (b) Interval Estimation
   (c) Point Estimation
      (i) Maximum Likelihood
      (ii)Robbins and Chapman procedure
9. Functions of random variables
   (a) Convolution
10. Higher moments
   (a) Moment generating function
   (b) Factorial moments
   (c) Pearson curves
11. Tschbycheff's inequality
12. Laws of large numbers
13. Central limit theorem
14. Law of Iterated Logarithm
15. Transformation of variables
   (a) Jacobian
   (b) Orthogonality

16. Distribution of linear forms: normal.
17. Distribution of sum of squares (χ^2) and s^2.
18. Comparing means for normal variates.
    (a) Normal deviate test
    (b) Student's t.
19. F-test for comparing two variances
20. Comparing group means (F): Analysis of variance
21. Use of χ^2 for contingency tables
    (a) Correction for continuity
    (b) Exact test
22. Variance stabilizing transformation
23. Combination of tests
24. Bivariate normal: correlation coefficient
    (a) Wishart distribution
    (b) Distribution of r; Fisher's z.
25. Neyman-Pearson Theory of Testing Hypotheses
    (a) Likelihood ratio test
26. Decision Theory
27. Non-central χ^2
28. Theory of Point Estimation
    (a) Cramer - Rao Theorem
    (b) Sufficiency
29. Order Statistics
30. Sequential analysis
31. General Linear Hypothesis
    (a) Vectors and Vector Spaces
    (b) General linear model with fixed effects
        (i) Estimation and error spaces
        (ii) Analysis of variance
    (c) Generalized t and F tests
    (d) Doubly classified material
    (e) Confidence Interval Estimation
    (f) Incomplete block designs: Balanced design
    (g) Analysis of covariance
    (h) Missing data
32. Variance component models

Table 3. Outline of Course in Basic Statistical Analysis

Perequisites: College algebra; a previous statistics course or graduate standing

Credits: Four hours each semester; two lecture hours and two two-hour laboratories.

Course Summary: This course introduces the student to the basic statistical tools for estimation and for testing hypotheses by use of empirical sampling from selected populations. These statistical tools are used to analyze actual experimental and survey data. Basic survey methods and experimental designs are discussed. This course is intended to parallel the Basic Theory course and is to be taken by Statistics majors or Ph.D minors, but is not intended as a service course for other departments.

Course outline

A. Introduction and Description of Populations

1. Introduction
   (a) Definitions
   (b) History
   (c) Present Uses of Statistical Analysis
   (d) Collecting data
   (e) Elementary Standards in Graphic Work
2. Attribute (classification) Data
   (a) Grouping data
   (b) Measures of association
3. Variables (Scaled) Data: Finite Populations
   (a) Frequency tables; class intervals; number of classes
   (b) Location measures (Central Tendency)
   (c) Dispersion measures
   (d) Skewness and Kurtosis
   (e) Combined Attribute and Variable Data
4. Some Special Hypothetical Populations:
   (a) Binomial and multinomial; Poisson; Normal; Uniform
   (b) Population parameters
5. Bivariate and multivariate continuous data

B. Sampling Experiments

1. Introduction
   (a) Universe and population ; statistics.
   (b) Use of samples: estimates and tests.
   (c) Criteria of estimates and tests.
   (d) Random sampling: with and without replacement
   (e) Surveys and experiments

2, Populations studied: normal, binomial, uniform and a special multimodal (acreage data).

3. Frequency tables and averages and standard errors for:
   (a) Mean, medians, midrange
   (b) Variance, range, mean deviation from median.
   (c) Normal deviates, t, χ^2,
   (d) Correlation and regression analysis

4. Confidence limits; one and two-tail tests; power.

5. Fitting distributions

6. Computed but no frequency tables: F; t-test for mean differences.

7. Enumeration Statistics
   (a) All probabilities given in advance
   (b) Goodness-of-fit when certain parameters are estimated from data
   (c) Contingency tables

C. Collection and Analysis of Data

1. Control charts

2. Non-parametric (Distribution Free) Methods: runs test; rank correlation; Wilcoxon rank sum test; sign test

3. Sequential Testing

4. Sample Surveys
   (a). Definitions; questionaire; collecting data
   (b) Simple random sampling - size of sample
   (c) Stratified random sampling - optimum allocation
   (d) Analytical surveys

5. Simple Regression

6. Multiple Regression
   (a) Abbreviated Doolittle technique
   (b) Orthogonal polynomials

7. Analysis of Variance for Designs in Complete Blocks: Randomized blocks and Latin squares.

8. Balanced Incomplete Blocks
   (a) Mention non-balanced designs

9. Factorial Experiments
   (a) Disproportionate frequency tables

10. Variance Components
   (a) All random: interactions and nested sampling
   (b) Mixed models: split-plot and stratified sampling.

Table 4. Computations for Samples from a Simulated Uniform Distribution, Using Data from Appendix Table 1 in Dixon and Massey.

($\mu = 49.5$, $\sigma^2 = 833.25$)

X	A	B	C	
1	10	98	91	
2	37	11	80	
3	08	83	44	
4	99	88	12	
5	12	99	63	
6	66	65	61	
7	31	80	15	
8	85	74	94	
9	63	69	42	
10	73	09	23	
SX	484	676	525	1685
$\bar{X}$	48.4	67.6	52.5	56.17
SX^2	33318	55102	35805	124225
$S(X+1)^2$	34296	56464	36865	127625
$SX^2 + 2SX + n$	34296	56464	36865	127625
Sx^2	9892.4	9404.4	8242.5	29584.2
$\chi^2_\sigma = Sx^2/\sigma^2$	11.87	11.28	9.89	35.50
s^2	1099.16	1044.93	915.83	1020.14
s	33.154	32.325	30.263	31.940
$s(\bar{X})$	10.48	10.22	9.57	5.83
$t(\mu) = (\bar{X}-49.5)/s(\bar{X})$	-.105	1.771	.313	1.14
UCL(μ)	67.61	86.34	70.04	66.07
LCL(μ)	29.19	48.86	34.96	46.26
UCL(σ^2)	2975	2828	2479	1670
LCL(σ^2)	585	556	487	695
Med. (M_e)	50	77	52.5	64
M.R.	53.5	54	53	53.5
Range(w)	91	90	82	91
S $X-M_e$	288	220	253	
C.V.	.685	.478	.576	.569
0, 1, 2	3, 3	2,2	3,5	8,10; 18
3,.., 9	7, 7	8,8	7,5	22,20; 42

Table 5. Frequency Tables for Various Statistics, based on 69 Samples of 10 each from Random 2 - Digit Numbers[1]

LCL[2]	$\bar{X}$	M_e	MR	w
20	0	1	0	0
25	0	4	1	0
30	6	3	0	0
35	3	9	4	1
40	15	10	5	1
45	12	11	24	0
50	14	9	24	0
55	8	4	10	2
60	6	6	0	2
65	3	3	0	7
70	2	4	0	2
75	0	3	1	11
80	0	2	0	11
85	0	0	0	14
90	0	0	0	13
95	0	0	0	5
Sf	69	69	69	69
Mean	49.70	50.22	49.64	80.84
Var.	97.47	209.76	44.56	153.05

(χ^2)[3]	f
1.5	1
2.5	1
3.5	0
4.5	3
5.5	9
6.5	7
7.5	11
8.5	4
9.5	12
10.5	7
11.5	6
12.5	3
13.5	2
14.5	1
15.5	2
Sf	69
Mean	8.63
Var.	8.50

t[3]	f
-2.25	2
-1.75	5
-1.25	4
-0.75	11
-0.25	13
0.25	15
0.75	7
1.25	6
1.75	5
2.25	1
2.75	1
6.25	1
Sf	69
Mean	.0905
Var.	1.754

$s\,\lvert X-M_e \rvert$[3]	f
74.5	1
104.5	1
134.5	4
164.5	9
194.5	15
224.5	17
254.5	11
284.5	9
314.5	2
Sf	69
Mean	216.24
Var.	2485.

[1] Taken from Appendix Table 1 of Dixon and Massey.

[2] LCL = lower class limit. Classes go from given LCL up to, but not including, next larger LCL.

[3] Middle of class interval

Table 6. Sampling Experiment with Enumeration Data

1. Assumed probabilities in a k-cell population

Cell	p_i	$20p_i$	n_i	R.N.
1	.2	4	n_1	0,1
2	.3	6	n_2	2,3,4
3	.2	4	n_3	5,6
4	.3	6	n_4	7,8,9
	1.0	20	n	

Draw samples of n = 20 from this population, giving n_1, n_2, n_3 and n_4 in each cell. Set up populations of random numbers (R.N.).

2. Compute for each sample $X_3^2 = \sum_{i=1}^{4} \frac{(n_i - np_i)^2}{np_i}$

3. Obtain distribution of X_3^2 and show it is approximately χ^2(3 df.). We will not determine power here.

4. Compute for each sample $X_1^2 = \frac{n\,(n_1n_4 - n_2n_3)^2}{(n_1+n_2)(n_3+n_4)(n_1+n_3)(n_2+n_4)}$

5. Obtain distribution of X_1^2 and show it is approximately χ^2(1 df.). Find 5% and 10% percentage points.

6. Compute X_{1c}^2 by use of Yates' correction for continuity. This is accomplished by using $(|n_1n_4 - n_2n_3| - \frac{n}{2})^2$ in numerator. If $|n_1n_4 - n_2n_3| < 1/2\ n$, set $X_1^2 = 0$. Check whether X_{1c}^2 is closer to χ^2 than X_1^2.

7. Compute $X_2^2 = \frac{n\,(n_1n_3 - n_2n_4)^2}{(n_1+n_2)(n_3+n_4)(n_1+n_4)(n_2+n_3)}$

(This is found by interchanging $n_3 + n_4$ in 4)

The percentage of X_2^2 values equal to a greater than the percentage points in 4 is the power of the test to detect a reversal of the probabilities.

8. Example, using the 20 digits in each of A, B, C (Table 9)

	n_1	n_2	n_3	n_4	X_3^2	X_1^2	X_2^2	X_{1c}^2
A	5	5	4	6	.42	.20	.20	0
B	4	2	3	11	7.08	3.78	.32	2.05
C	5	9	3	3	3.50	.36	.36	.01
Total	14	16	10	20	1.11	1.11	2.45	.62