Political Science Course Syllabi Collection

METHODOLOGY

edited by

John R. Freeman &
W. Phillips Shively
University of Minnesota

DEVELOPED BY

The American Political Science Association

WITH SUPPORT FROM

The Fund for the Improvement
of Postsecondary Education

1992

Elementary Statistics for Political Research

Kenneth Janda, Northwestern University*

This course is offered on a quarter basis for 10 weeks. The average number of students ranges from 90 to 100, typically half of whom are majors. The course is required for Political Science and Sociology graduate students, and satisfies a requirement for undergraduate majors as well. There is one lecture and one discussion session each week.

*Kenneth Janda is a Professor who has been teaching for 31 years.

Political Science C10: *Elementary Statistics for Political Research*
Sociology C30: *Basic Statistics for Sociology*

Fall, 1991 **Mr. Janda**

This course in statistics is designed to integrate research methods with the substance of political and social research. Statistical techniques and applications have been chosen for their relevance to political science and sociology, fields that many people do not regard as quantitative. This approach should make the study of statistics more meaningful and valuable to those who are uncertain about the role of statistical analysis in the softer social sciences.

In addition to its focus on social and political research, this course differs from most other statistics courses in using the SPSS computer program as an integral part of statistical training. Computers are now important and commonplace tools of political research, and it is vital for the research worker to know how to use standard computer programs for statistical analysis. Each student will be expected to record a small set of data for use in statistical exercises with the computer and to perform other analyses on data called from computer storage. Because the class is large, we will use SPSS on the IBM 4381 mainframe computer, accessed through remote terminals. Students who have their own microcomputers will be instructed about statistical programs for IBM-compatible and Macintosh computers.

REQUIRED TEXT

Kirk, Roger E. *Statistics: An Introduction, Third Edition* (Fort Worth: Holt, Rinehart and Winston, 1990)

RECOMMENDED TEXT

Marija J. Norusis, *SPSS Introductory Statistics Student Guide* (Chicago: SPSS Inc., 1990)

These books can be purchased at Norris Center or at SBX. There will be additional reading outside of these texts on reserve at the Library and available for purchase at CopyCat. They illustrate applications of the statistics studied, and they also should provide ideas for your research paper.

TEACHING METHODS

The course will be conducted in lectures on Monday, Tuesday, Wednesday, and Friday at 12:00 noon in Swift 107, where we have special video equipment to display computer output via a terminal linked to the campus IBM computer. Optional discussion sections (I will hold them at *my* option and you can attend them at your option) will be held on Thursday. There is just too much material to be learned in one quarter to cover it well in only three lectures per week. While you are asked to attend lectures an extra hour a week without increased course credit, remember that I am teaching the extra lecture without increased pay. I would not impose the burden on us if it were not needed to cover statistics properly in one quarter.

EVALUATION OF PERFORMANCE

Your performance in this course will be evaluated through a 1/3 term exam (worth 15% of the final grade), a 2/3 term exam (worth 25%), a final examination (worth 35%), and a short paper (worth 25%). Your scores will be cumulated to produce a total point score on which your grade will be based. The undergraduates' scores will set the curve for the graduate students. Because this course stresses effective use of the computer, your grade may suffer if you are *grossly inefficient in using computing funds. Each student will be* awarded a sum of computing funds at the start of the course. Additional funds will be made available as needed to complete your assignments.

TEACHING ASSISTANTS

I will be assisted by two or more graduate student who will hold weekly office hours to help with your computer runs and answer questions about the lectures. The TAs will also grade your research paper.

RESEARCH PAPER

The paper is intended to demonstrate your competence in applying statistics to a specific research problem of your own formulation. You will be expected to devise a concise and non-trivial hypothesis that can be tested with data available for computer analysis. You must state briefly the source of your research problem or question, formulate it as a hypothesis for testing with available data (more on this later), execute the appropriate test, and draw a conclusion about the truth or falsity of the hypothesis-- within a maximum of 5 typewritten pages, double-spaced. Evaluation of this exercise will be based mainly on clarity of presentation and statistical craftsmanship rather than on the substantive or theoretical importance of the problem.

The expected form and style of the paper can be likened to the "research notes" that are occasionally published in political science journals (especially in recent issues of the *Journal of Politics*). We will be reading several examples of such articles. Your 5-page paper will not be due until **December 3** but a one-page progress report stating the hypothesis and data set is due on **Monday, November 18.**

READINGS AND ASSIGNMENTS

The outline of topics on the next page constitutes a table of contents to the course. **YOU SHOULD PURCHASE A LOOSELEAF NOTEBOOK** to contain your lecture notes and the numerous handouts that you will receive in class. Unless you have some way to organize that material for effective retrieval later, you will be left with a confusing pile of papers.

Reading assignments from the required texts (and from material placed on Reserve) are specified on subsequent pages in the syllabus. Be certain to read the accompanying "comments" on the assignments, for they will direct your attention to what you are expect to learn.

Course Topics

I. QUANTITATIVE ANALYSIS IN POLITICAL RESEARCH

Sept.23: Inspirational Message: Why You Should Take Statistics
Sept.24: Statistics and the Logic of Inquiry
Sept.25: Why All Analysis Is Quantitative: Measurement Theory

II. USING COMPUTERS IN QUANTITATIVE ANALYSIS

Sept.27: Learning to Use Remote Terminals
Sept.30: Recording Data in Machine-Readable Form
Oct.1: Creating an SPSSx Data File
Oct.2: A Second SPSSx Run
Oct.3: Optional Lab Session on the Computer if Needed
Oct4.: Accessing an Existing Data Set: The 51 States Data

III. DESCRIBING A SINGLE VARIABLE: UNIVARIATE STATISTICS

Oct.7: Frequency Distributions
Oct.8: Measures of Central Tendency and Dispersion
Oct.9: Measures of Dispersion
Oct.11: The Normal Distribution

Oct.14: *1/3 EXAMINATION*

IV. MEASURING RELATIONS BETWEEN TWO CONTINUOUS VARIABLES

Oct.15: The Product-Moment Correlation
Oct.16: Interpreting the Correlation Coefficient
Oct.18: Linear Regression

V. MEASURING RELATIONS BETWEEN TWO DISCRETE VARIABLES

Oct.21: Contingency Tables or Cross-Tabulations
Oct.22: Measures of Association for Nominal and Ordinal Data--Take Your Pick

VI. STATISTICAL INFERENCE: ELEMENTARY PROBABILITY AND SAMPLING THEORY

Oct.23: A Priori and Empirical Probabilities
Oct.25: Discrete and Continuous Probability Distributions
Oct.28: Difference of Means Test: Single Sample
Oct.29: One-Tailed and Two-Tailed Tests
Oct.30: Significance Tests for Correlation Coefficients

Nov.1: *2/3 EXAMINATION*
Nov.4: Discussion of research papers

VII. TESTING HYPOTHESES AND PREDICTING TO POPULATIONS

Nov.5: Difference of Means Test: Two Samples
Nov.6: The t-Test in Political Research

VIII. ANALYSIS OF VARIANCE

Nov.8: *One-Way Analysis of Variance*
Research progress reports due!
Nov.11: One-Way Analysis of Variance in Research
Nov.12: Eta-*Squared and* ANOVA *Assumptions*
Nov.13: Interval Estimation

IX. MULTIPLE REGRESSION

Nov.15: Multiple Regression
Nov.18: More on Multiple Regression
Nov.19: Multiple Regression in Research
Nov.20: More Regression Analysis and Dummy Variables
Nov.22: Comparing Regression with Analysis of Variance

X. INTRODUCTION TO THE WORLD OF NON-PARAMETRIC TESTS

Nov.25: Non-Parametric Tests: Chi-Square

Nov.26: Help on Research Papers
Nov.27: Help on Research Papers (continued)

XI. REVIEW FOR FINAL EXAMINATION

Dec.2: Review -- Theory, Measurement, and Univariate Statistics
Dec3: Review -- Statistical Inference and Hypothesis Testing
RESEARCH PAPERS DUE
Dec.4: Review -- Bivariate Distributions: Strength, Form, Significance
Dec.6: Review -- Multivariate Analysis

June 15: **Final Examination** at 12:00

I. QUANTITATIVE ANALYSIS IN POLITICAL RESEARCH

SEPTEMBER 23:
INSPIRATIONAL
MESSAGE

This will be the first meeting of class. In place of reading from the texts, there will be a moving sermon. The topic is "Why You Should Take Statistics."

SEPTEMBER.24:
STATISTICS AND THE
LOGIC OF
INQUIRY

Kirk, Ch.1: "Introduction to Statistics," *only* 1-10

Kirk draws the important distinction between descriptive statistics and inferential statistics. The first part of the course will treat descriptive statistics; the second part covers inferential. This order is the reverse of most statistics courses, which cover probability theory and related topics necessary to statistical inference at the very beginning. I think it wiser pedagogically to begin with the simpler descriptive statistics, for which the research applications are more readily illustrated.

SEPTEMBER 25:
WHY ALL
ANALYSIS
QUANTITATIVE:
MEASUREMENT
THEORY

Kirk, Ch.1: "Introduction to Statistics," 10-26

The philosopher, Alfred North Whitehead, put it this way:

> *Through and through the world is infested with quantity: To talk sense is to talk quantities. It is no use saying the nation is large -- How large? It is no use saying that radium is scarce -- How scarce? You cannot evade quantity. You may fly to poetry and music, and quantity and number will face you in your rhythms and your octaves.*

Kirk distinguishes among several types of data--qualitative v. quantitative, discrete v. continuous--and several "levels" of measurement: nominal, ordinal, interval, and ratio. These are standard classification schemes, and you should learn them well: Relying on nominal-level measurement as a type of quantification, I will argue in class that *all* social analysis is quantitative. If you think otherwise, prepare your argument and be ready to dispute my position.

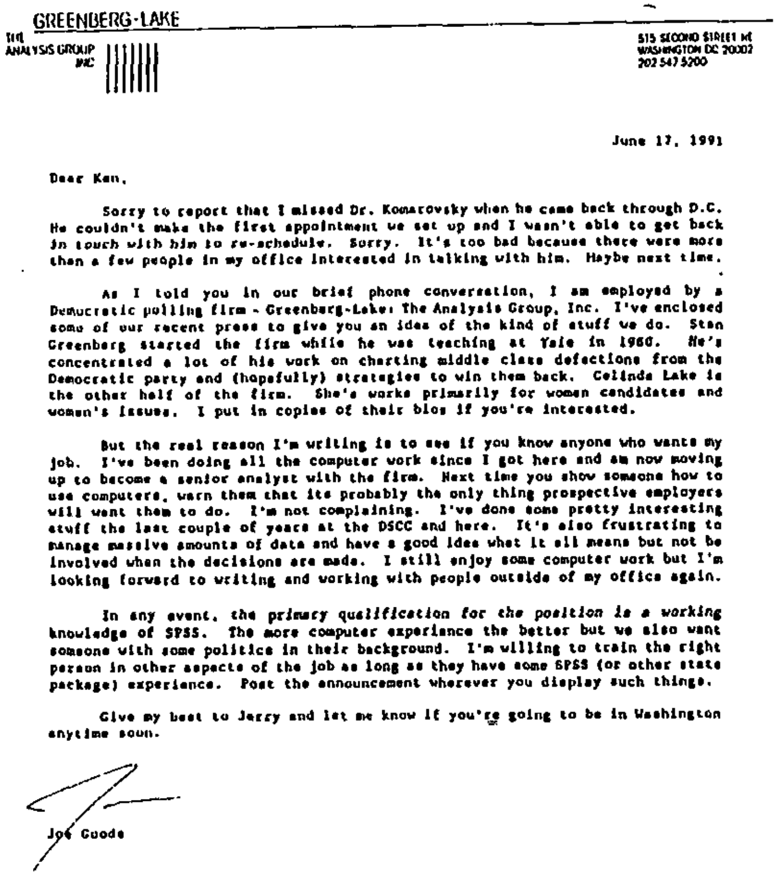
Assignment: Exercises 11 (a, b, d, g); 15, and 21 in Kirk. Note that the ANSWERS are at the back of the book. The exercises will not be submitted or graded, but they will serve as models for some items on the various examinations. If you do the exercises regularly, you should find the exams more familiar.

Letters from former students who took this statistics course:

If you are on the Left
(Greenberg-Lake does polls for Democratic candidates)

You *need* to know statistics to make sense out of politics!

If you are on the Right
(Market Strategies does polls for Republican candidates)



GREENBERG-LAKE
THE ANALYSIS GROUP INC

515 SECOND STREET NE
WASHINGTON DC 20002
202 547 5200

June 17, 1991

Dear Ken,

Sorry to report that I missed Dr. Komarovsky when he came back through D.C. He couldn't make the first appointment we set up and I wasn't able to get back in touch with him to re-schedule. Sorry. It's too bad because there were more than a few people in my office interested in talking with him. Maybe next time.

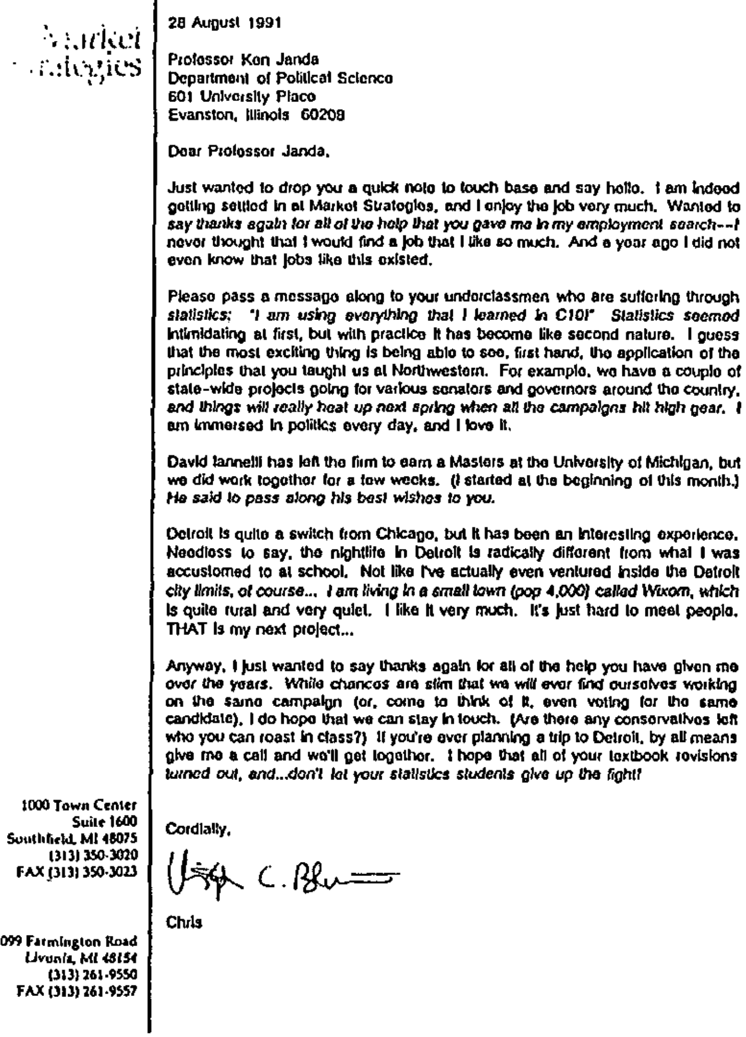
As I told you in our brief phone conversation, I am employed by a Democratic polling firm - Greenberg-Lake: The Analysis Group, Inc. I've enclosed some of our recent press to give you an idea of the kind of stuff we do. Stan Greenberg started the firm while he was teaching at Yale in 1980. He's concentrated a lot of his work on charting middle class defections from the Democratic party and (hopefully) strategies to win them back. Celinda Lake is the other half of the firm. She's works primarily for women candidates and women's issues. I put in copies of their bios if you're interested.

But the real reason I'm writing is to see if you know anyone who wants my job. I've been doing all the computer work since I got here and am now moving up to become a senior analyst with the firm. Next time you show someone how to use computers, warn them that its probably the only thing prospective employers will want them to do. I'm not complaining. I've done some pretty interesting stuff the last couple of years at the DSCC and here. It's also frustrating to manage massive amounts of data and have a good idea what it all means but not be involved when the decisions are made. I still enjoy some computer work but I'm looking forward to writing and working with people outside of my office again.

In any event, the *primary qualification for the position is a working* knowledge of SPSS. The more computer experience the better but we also want someone with some politics in their background. I'm willing to train the right person in other aspects of the job as long as they have some SPSS (or other stats package) experience. Post the announcement wherever you display such things.

Give my best to Jerry and let me know if you're going to be in Washington anytime soon.

Joe Goode



Market Strategies

28 August 1991

Professor Ken Janda
Department of Political Science
601 University Place
Evanston, Illinois 60208

Dear Professor Janda,

Just wanted to drop you a quick note to touch base and say hello. I am indeed getting settled in at Market Strategies, and I enjoy the job very much. Wanted to *say thanks again for all of the help that you gave me in my employment search--I* never thought that I would find a job that I like so much. And a year ago I did not even know that jobs like this existed.

Please pass a message along to your underclassmen who are suffering through *statistics: "I am using everything that I learned in C10!" Statistics seemed* intimidating at first, but with practice it has become like second nature. I guess that the most exciting thing is being able to see, first hand, the application of the principles that you taught us at Northwestern. For example, we have a couple of state-wide projects going for various senators and governors around the country, *and things will really heat up next spring when all the campaigns hit high gear. I* am immersed in politics every day, and I love it.

David Iannelli has left the firm to earn a Masters at the University of Michigan, but we did work together for a few weeks. (I started at the beginning of this month.) *He said to pass along his best wishes to you.*

Detroit is quite a switch from Chicago, but it has been an interesting experience. Needless to say, the nightlife in Detroit is radically different from what I was accustomed to at school. Not like I've actually even ventured inside the Detroit *city limits, of course... I am living in a small town (pop 4,000) called Wixom*, which is quite rural and very quiet. I like it very much. It's just hard to meet people. THAT is my next project...

Anyway, I just wanted to say thanks again for all of the help you have given me *over the years. While chances are slim that we will ever find ourselves working* on the same campaign (or, come to think of it, even voting for the same candidate), I do hope that we can stay in touch. (Are there any conservatives left who you can roast in class?) If you're ever planning a trip to Detroit, by all means give me a call and we'll get together. I hope that all of your textbook revisions *turned out, and...don't let your statistics students give up the fight!*

Cordially,

C. Bl

Chris

1000 Town Center
Suite 1600
Southfield, MI 48075
(313) 350-3020
FAX (313) 350-3023

14099 Farmington Road
Livonia, MI 48154
(313) 261-9550
FAX (313) 261-9557

II. USING COMPUTERS IN QUANTITATIVE ANALYSIS

SEPTEMBER 26:
OPTIONAL LAB:
COMPUTING
FACILITIES TOUR

Meet as usual in Swift 107 at 9:00, and we will tour the facilities and get a start on Friday's assignment. It will help if you do the reading in advance.

SEPTEMBER 27:
LEARNING TO
USE REMOTE
TERMINALS

Kirk, Ch.2: . . . . "Computer Supplement," *only* 61-69
VM/CMS Handbook, Ch.1: "Introduction"
Ch.2: "Logging On VM/CMS"
Ch.3: "How to Use Xedit"
Ch.4: "How to Use CMS"

Although there will be some computer wizards in the class, many of you will be new to the mysteries of computer analysis. Our work in this class can be done using remote terminals to access the IBM 4381 computer, which is the "mainframe" operated by Northwestern's Academic Computing and Network Services (ACNS). The Vogelback building has about 10 "public" remote terminals available for your assignments, but they become heavily used as the quarter progresses. More terminals connected to the IBM are available in the Library, located in the hallway to the Periodicals Room. If you have acess to your own computer, you can access the mainframe over telephone lines using a device called a modem. If you have a Macintosh computer, you can copy a modem program called "MacKermit" at Vogelback to your own disk. If you have an IBM-compatible computer, bring a disk to Vogelback and ask for a copy of PIBTERM.

Assignment: Guided by Kirk's explanation of how to use a table of random numbers on page 264, refer to his Table D.1 of Random Numbers on pages 650-651. Then select 10 states at random from the list of data on American states on the next page. IT IS IMPORTANT TO CHOOSE YOUR TEN CASES *EXACTLY* AS KIRK EXPLAINS. Don't choose your "favorite" states or use your own idea of "random" sampling. Underline or "highlight" the ten states that constitute your random sample. In class, I will describe how you will use a VDT (video display terminal) to record data on a sample of American states for later computer analysis.

SEPTEMBER 30:
RECORDING
DATA IN MACHINE-
READABLE FORM

Assignment: Enter the data on your random set of 10 states into a computer file file at the terminal using **XEDIT** (the IBM Full Screen Editor) and a file that you should save as **MYDATA**. Use one line for each state, following the instructions in the handout supplied for this purpose. Then print the contents of the file on the line printer as instructed. The results should help you understand the relationship between the VDT display of a computer file and a printed version of that file. Bring the printout to class as evidence of your triumph over the machine. (If the machine wins the first round, submit the output later.)

NAME	BLACKPOP	PCTBLACK	REAGAN84	BUSH88	VOTES80	VOTES90
Columns 1-8	9 - 14	15-18	19-20	21-24	25-26	27-28
ALABAMA	996335	26.0	61	59.2	9	9
ALASKA	13643	3.4	67	59.7	3	3
ARIZONA	74977	2.8	66	60.0	7	8
ARKANSAS	373768	16.0	61	56.4	6	6
CALIFORN	1819281	7.7	58	51.1	47	54
COLORADO	101703	3.5	63	53.1	8	8
CONNECTI	217433	7.0	61	52.0	8	8
DELAWARE	95845	16.0	60	55.9	3	3
DIST OF	448906	70.0	13	14.3	3	3
FLORIDA	1342688	14.0	65	60.9	21	25
GEORGIA	1465181	27.0	60	59.7	12	13
HAWAII	17364	1.8	55	44.7	4	4
IDAHO	2716	.3	73	62.1	4	4
ILLINOIS	1675398	15.0	56	50.7	24	22
INDIANA	414785	7.6	61	59.8	12	12
IOWA	41700	1.4	54	44.5	8	7
KANSAS	126127	5.3	67	55.8	7	6
KENTUCKY	259477	7.1	60	55.5	9	8
LOUISIAN	1238241	29.0	61	54.3	10	9
MAINE	3128	.3	61	55.3	4	4
MARYLAND	958150	23.0	52	51.1	10	10
MASSACHU	221279	3.9	51	45.4	13	12
MICHIGAN	1199023	13.0	59	53.5	20	18
MINNESOT	53344	1.3	50	45.9	10	10
MISSISSI	887206	35.0	62	59.9	7	7
MISSOURI	514276	10.0	60	51.8	11	11
MONTANA	1786	.2	60	52.1	4	3
NEBRASKA	48390	3.1	71	60.1	5	5
NEVADA	50999	6.4	67	58.9	4	4
NEW HAMP	3990	.4	69	62.5	4	4
NEW JERS	925066	13.0	60	56.2	16	15
NEW MEXI	24020	1.8	60	51.9	5	5
NEW YORK	2402006	14.0	54	47.5	36	33
NORTH CA	1318857	22.0	62	58.0	13	14
NORTH DA	2568	.4	66	56.0	3	3
OHIO	1076748	10.0	59	55.0	23	21
OKLAHOMA	204674	6.8	69	57.9	8	8
OREGON	37060	1.4	56	46.6	7	7
PENNSYLV	1046810	8.8	54	50.7	25	23
RHODE IS	27584	2.9	52	43.9	4	4
SOUTH CA	948623	30.0	64	61.5	8	8
SOUTH DA	2144	.3	63	52.9	3	3
TENNESSE	725942	16.0	59	57.9	11	11
TEXAS	1710175	12.0	64	56.0	29	32
UTAH	9225	.6	75	66.2	5	5
VERMONT	1135	.2	58	51.1	3	3
VIRGINIA	1008668	19.0	63	59.7	12	13
WASHINGT	105574	2.6	56	48.5	10	11
WEST VIR	65051	3.3	55	47.5	6	5
WISCONSI	182592	3.9	55	47.8	11	11
WYOMING	3364	.7	70	60.5	3	3

OCTOBER 1:
CREATING AN SPSS DATA FILE

For those of you with microcomputers, there is also an IBM-PC version called SPSS/PC+, and a Macintosh version of SPSS. However, both are expensive and both require a great deal of free disk space. Another version, called SPSS/PC+ Studentware, is available for only $40, but it runs only on IBM-compatible computers and it still requires a hard disk even though it is limited to 50 variables. Studentware must be ordered through a bookstore. Please give me your name if you are intrested, and I will coordinate the orders.

Kirk, Ch.2: . . . . "Computer Supplement," 69-79
Norusis, Ch.1: "Getting Started: A few Useful Terms," 1-5
Ch.2: "Preparing and Defining Data for SPSS Analysis,"*only* 6-16
Ch.6: "Listing Cases: Procedure LIST," 54-57

SPSS is the most widely-used system of computer programs for social analysis. It was conceived by a humble graduate student in political science at Stanford University in the late 1960s. Norman Nie, now President of SPSS Inc. and Professor of Political Science at the University of Chicago, has seen his creation spread to thousands of universities and businesses across the world. SPSS has gone through numerous updates, and we will be using what is curiously called "Release 4"--despite the fact that there have been at least 10 other editions.The first step in using the computer for data analysis with SPSS is to prepare the "data definition" commands that precede your machine-readable data. Norusis starts you off by describing the SPSS language and explaining how to set up a data file for computer analysis.

SPSS has so many capabilities that even skilled researchers never learn the system completely. In this course, you will only learn its basic statistical procedures. These will be enough to equip you for elementary data analysis. Learning a computer system resembles learning a new language, and you will simply have to spend some effort in learning new terms and computer grammar. Out of mercy, I decided not to require the definitive *SPSS Reference Guide*, which is 800 pages of intimidation. Instead, we are using the much shorter *SPSS Introductory Statistics Student Guide* by Marija Norusis. Be properly thankful.

Assignment: Using a remote terminal, enter the IBM and recall your 10-state data set following the instructions passed out in class. Precede your data with the control cards necessary to create an SPSS file for your ten states. Create an SPSS file and use the LIST procedures to list your ten cases for these variables: STATE BUSH88 VOTES80 VOTES90.

Also do these Exercises in Norusis: Ch.1: 1, 2, 3; Ch.2: 1-2.

OCTOBER 2:
A SECOND SPSS RUN

Norusis, Ch.3: "Data Transformation and Selection," 20-35

As described in these chapters by Norusis, SPSS is not only a collection of programs for statistical analysis; it is also a general "data management" system that you can use to recode existing values, to create new variables, and to select portions of your data set for analysis.

Assignment: Make another SPSS run with your small data file, but use the COMPUTE command in SPSS to form the new variable, VOTEGAIN, by subtracting VOTES80 from VOTES90. Once again, run the LIST procedure, but this time list these variables: STATE VOTES90 VOTES80 VOTEGAIN. Check to see that you have computed VOTEGAIN correctly.

Do these Exercises in Norusis, Ch.3: 1-3

OCTOBER 3:
OPTIONAL LAB

This session will only deal with computer usage. If you feel comfortable using the computer by now, you need not attend this session.

OCTOBER 4: ACCESSING AN EXISTING DATA SET: THE 51 STATES DATA

Norusis, Ch.4: "*SPSS System Files and File Management*," 36-48
Ch.5: "SPSS Session Control," 50-53 (You will need to know *only* these commands: GET, FILE, SAVE, OUTFILE, DISPLAY, SORT CASES.)
VM/CMS Handbook, "Running SPSSX on VM/CMS with FILELIST"

You do not have to create an SPSS file every time you analyze data. Instead, the file can be "saved" with a SAVE command and then retrieved with a GET command. The full data set for 51 states (all states plus the District of Columbia) and numerous variables has already been made into an SPSS file and saved for your usage under the file name, **STATES**. We will use this full data set in the assignment below. You should understand that STATES was created in a process akin to the one that you used to make your small data set. If you wish to make another data set sometime for analysis in another course, you should be able to reconstruct the process. For the rest of this course, however, we will be using existing SPSS data sets.

Assignment: Enter XEDIT and prepare the SPSS commands that will select the STATES data for analysis and print the labels of the variables in the STATES file. Choose the option to enter SPSS commands, and run SPSS by entering these three commands on separate lines:

```
GET FILE STATES
DISPLAY LABELS
LIST VARIABLES=STATE BUSH VOTEGAIN
```

Then exit XEDIT and type SPSSX STATES to run the job. The output from this run will be stored in a file called STATES LISTING. Print the output using the command GPRINT STATES LISTING and save it for reference in later assignments.

(As described in Norusis, the SPSS command DISPLAY and the subcommand LABELS gives a description of the LABELS of all the variables on the saved file.)

III. DESCRIBING A SINGLE VARIABLE: UNIVARIATE STATISTICS

OCTOBER 7:
FREQUENCY DISTRIBUTIONS

Kirk, Ch.2: "Frequency Distributions and Graphs," 29-61 and 69-79
Norusis, Ch.7: "Data Tabulation: Procedure FREQUENCIES," 59-73

In our American government text, *The Challenge of Democracy*, my co-authors and I rely on graphs to display data that are usually put in tables. We feel that well-constructed graphs communicate points more memorably than tables.

Assignment: Run FREQUENCIES for VOTES80 VOTES90 and VOTEGAIN. You will have to study Norusis Ch.7 to learn how to do this. Include the BARCHART subcommand.
Exercises in Norusis, Ch.7: Syntax 1 and Statistical Concepts 1, 2, and 3.
Kirk, Ch.2: 5 (a,b,d), 59 (a,b)

OCTOBER 8:
MEASURES OF CENTRAL TENDENCY

Kirk, Ch.3: "Measures of Central Tendency," *only* 81-106

Frequency distributions are informative but clumsy. The essential information in frequency distributions can often be summarized with two types of statistics: (1) *measures of central tendency* and (2) *measures of dispersion* or *variation*. Each type of statistic tells something about the nature of the full distribution. We begin with the simpler summary statistics: measures of central tendency.

Assignment: Using the STATES data file and variables on PCTWOMEN, PCTBLACK, and BUSH88, run FREQUENCIES. Because these are continuous variables, you will not want to print the frequency counts, so enter the following subcommand after the FREQUENCIES= command:

/FORMAT=NOTABLE *(this reads "No Table" not "notable")*

Also include a STATISTICS subcommand that asks for MEAN, MEDIAN, MODE, STDDEV, VARIANCE, KURTOSIS, SKEWNESS, MINIMUM, MAXIMUM, and RANGE. Study the output and be sure you know what you have computed -- for the mean, median, and mode. We will study the other statistics the next session.

Do exercises in Kirk, Ch.3: 25 (a-f) and 28(a, e, f)

OCTOBER 9:
MEASURES OF DISPERSION

Kirk, Ch.4: "Measures of Dispersion, Skewness, and Kurtosis," 113-151
Norusis, Ch.8: "Descriptive Statistics: DESCRIPTIVES," 74-90
Ch.9: "Looking First: Procedure EXAMINE," 92-110

Variability in observations lies at the heart of statistical analysis *and* social analysis. Most research seeks to understand why some person, group, or nation differs on some variable from other persons, groups, or nations. Indeed, if the cases did not differ, then we would not be observing a *variable*. Fundamental to variability in statistical analysis is the concept of *variance*. Be certain you know it well-- along with the related concept, *standard deviation*. This is definitely knowledge to be tested on examinations.

Assignment: Compare the variables you ran for yesterday's FREQUENCIES assignment for "skewness" and "kurtosis." Which variable is closer to a normal distribution? Your class lecture notes show how nonlinear transformations can be used to "normalize" skewed distributions.

Do exercises in Kirk, Ch.4: 9, 21 (a)

OCTOBER 11:
THE NORMAL DISTRIBUTION

Kirk, Ch. 9: "Normal Distribution and Sampling Distributions," *only* 283-297

Thorough understanding of the normal distribution is very important to inferential statistics in the second part of the course. If you can do the problems and exercise assigned below, you should have no trouble later.

Assignment: Do exercises in Kirk, Ch.9: 3 (a-c), 4 (a-d), 5 (a-c), 6 (a-c), 7 (a-c), 19, and 20 (a-c).

OCTOBER 14:
1/3 Examination

This examination is worth 15% of the course grade. It will cover everything up to this point, including computer concepts and commands. It will be both multiple-choice and "fill in the blank." You may bring a calculator.

Sports and z-scores: interpreting Mike O'Donnell's *Chicago Tribune* article of 8/11/85

With Ozzie around, It Paid for Cards to Let Sutter Go

A few years ago, the St. Louis Cardinals (a baseball team), allowed their top relief pitcher, Bruce Sutter, to go to another team by failing to match the other team's salary offer. However, the Cardinals then gave even more money to keep their shortstop, Ozzie Smith. *Chicago Tritune* Sports reporter Michael O'Donnell sought to compare the value of Ozzie Smith, at his position compared to Sutter at his. O'Donnell used z-scores for the purpose.

The performances of 21 shortstops were scored on six indicators:

1. errors per 100 chances
2. total chances per game
3. double plays per game
4. batting average
5. on-base percentage
6. runs created

The performances of 50 pitchers were scored on six indicators:

1. wins
2. earned run average
3. innings pitched
4. baserunners per nine innings
5. strikeouts per nine innings
6. saves

Means and standard deviations were computed for each of the 12 indicators or "variables," with N = 21 for one set and N= 50 for the other.

Smith's performance on each variable was transformed into z-scores.

The difference between Smith's "errors per 100 chances" and the mean for all 21 was divided by the standard deviation of the distribution.

The sign of the resulting z-score showed whether Smith ranked above or below the mean, and its value shows how far -- in standard deviation units.

Sutter's performances were also transformed into z-scores. O'Donnell then SUMMED the 6 z-scores separately for Smith and Sutter to produce two summed z-scores:
Sutter: 6.44
Smith: 5.21

By itself, this analysis shows that, RELATIVELY SPEAKING, Sutter ranked higher among relief pitchers than Smith did among shortstops. However, O'Donnell then looked at other statistics --

- the fact that Smith played a full nine innings a game, not just the last two when the team was ahead, which is common for a relief pitcher
- these statistics influence his final judgment that Smith was more valuable than Sutter.

So really, O'Donnell's z-score analysis did not decide for him which player was more valuable; it was only one factor in his analysis. The article was REALLY written so that I could use it in class. Thanks, Michael O'Donnell.

IV. MEASURING RELATIONSHIPS BETWEEN TWO CONTINUOUS VARIABLES

OCTOBER 15: THE PRODUCT-MOMENT CORRELATION

Kirk, Ch.5: "Correlation," *only* 155-170

If your variables are continuous, or if you can treat them as points along a conceptual continuum, relationships can be measured and expressed precisely and concisely through the twin techniques of product-moment correlation and regression. Correlational analysis is one of the most common techniques in social research. In essence, it tells us in what direction two variables are related and how *strongly* they are related.

Assignment: Do exercises in Kirk, Ch.5: 15

OCTOBER 16: INTERPRETING THE CORRELATION COEFFICIENT

Kirk, Ch.5: "Correlation," 171-198
Norusis, Ch.14: "Measuring Linear Association: CORRELATIONS," 182-197
Marilyn J. Field, "Determinants of Abortion Policy in the Developed Nations," *Policy Studies Journal*, (Summer, 1979), 771-781. (On Reserve)

The strength of a association between two interval-level variables is actually expressed by the *coefficient of determination*, which is simply the square of the product moment correlation. Field tests several hypotheses about the policies of different nations toward abortion, including one holding that Catholic countries would have more conservative abortion policies than non-Catholic ones. Are these hypotheses supported or refuted by her data analysis?

Assignment: Formulate a hypothesis about the effect of socioeconomic characteristics on politics in American states and test it using data from the STATES file. You must select a dependent variable (the political outcome you wish to explain) and an independent variable (the socioeconomic characteristic that is a likely explanation of the outcome). You should determine for yourself whether the hypothesis was supported by the data. Print the results obtained from running CORRELATION. Ask for STATISTICS ALL -- which generates mean, standard deviations, and cross-product deviations and covariances between all pairs of variables. Using the statistics generated from the XY cross-product deviations and the X and Y variances, try to compute the correlation value, *r*, as explained (but not too clearly) in Kirk, pp.165-166. Try to work this out on your own before I show how in class. **A QUESTION LIKE THIS WILL BE ON THE 2/3 EXAM.**

Do exercises in Kirk, Ch.5: 22 (a), 24 (a, c), 29 (a,d,e,g)

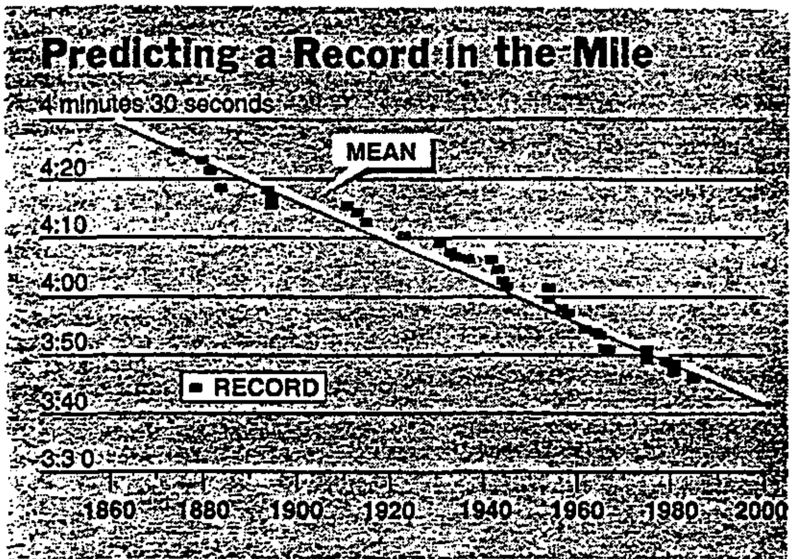
OCTOBER 18: LINEAR REGRESSION

Kirk, Ch.7: "Regression," *only* 201-217 and 224-225
Norusis, Ch.13: "Plotting Data: PLOT," 166-180
(You are resposible only for these subcommands: PLOT, VERTICAL/ HORIZONTAL, FORMAT=REGRESSION, HSIZE/VSIZE, and MISSING.)

Whereas correlational analysis tells the *strength* of a relationship, regression analysis indicates the specific *form* of a relationship, providing a capacity for predicting to Y with knowledge of X. A great deal of advanced statistical analysis is predicated on the ideas in simple linear regression. Learn this well.

Assignment: Run the PLOT procedure with the subcommand PLOT=[the variables used above with CORRELATION]. Also use the subcommand: FORMAT=REGRESSION. Compare the correlation coefficient given on your PLOT printout with that on your CORRELATION printout. Are they the same? Return to your PLOT output and interpret these other statistics: INTERCEPT, SLOPE, and S.E. OF EST that are associated with the regression analysis. The additional statistics computed for PLOT will be interpreted later.

Do exercises in Kirk, Ch. 6: 1, 5



Predicting a Record in the Mile

New mathematics textbooks try to interest students with intriguing real-life problems. They forego the traditional emphasis on mechanical drills and focus instead on comprehension, relying on calculators to handle the numbers.

This example, from an 11th-grade textbook developed by the University of Chicago School Mathematics Project, uses the history of record times in the men's mile to teach students about graphs and the limits of probability. The chart plots world records over time and suggests a formula for predicting records. Using the formula, a student could predict a probable world record in the year 2000 of 3:42.16, a plausible dip from the current record of 3:46.32 held by Steve Cram of Britain.

The lesson also attempts to show that while mathematics relies on logic and patterns, finding answers takes more than just plugging in numbers. Thus, students are asked to extend the line of records to the year 3000. Such a graph would suggest that sometime around that year, a man would finish the mile-run in no time, and the year after that in negative time.

V. MEASURING RELATIONSHIPS BETWEEN TWO DISCRETE VARIABLES

OCTOBER 21: CONTINGENCY TABLES OR CROSS-TABULATIONS

Norusis, Ch.10: "Crosstabulation . . .CROSSTABS," *only* 112-116 and 129-137 (You will only be responsible for the TABLES and CELLS subcommands. We will take up the STATISTICS subcommand later.)

Unlike economists, who like to think they deal in "real" numbers, political scientists and sociologists often must rely on nominal and ordinal data. These data are commonly analyzed in contingency tables, often only with the aid of percentage comparisons. These readings explain the time-honored practice of percentage analysis of data in tables.

Assignment: A national survey of voters before and after the 1988 election has been prepared as an SPSS file and stored as file "VOTE88." (You can get a list of all the variables in the VOTE88 file by typing GETINFO at the Ready prompt in CMS. Then respond to the menu choices. Print the variable list for later use.) Use the CROSSTABS procedure with this file to crosstabulate the respondent's vote for president in 1988 *separately* by the respondent's party identification, then by sex. (Find out the SPSS variable names yourself from the variable listing.) In the same run, make another table crosstabulating vote BY party BY sex. (Using three "BY" commands in this way will control for sex.) Specify the correct CELLS subcommand to compute percentages by *column* and the correct STATISTICS subcommands to compute LAMBDA, BTAU, CTAU, D, GAMMA, and CORR. (We will discuss these tomorrow; ignore them today.) Concentrate on interpreting the cell entries in the tables. Does controlling for sex make a difference in the relationship between party identification and vote? That is, was there a "gender gap" between Bush and women voters?

Do exercises in Norusis, Ch.10: Syntax 1 and 2; Statistical Concepts 6 and 7.

OCTOBER 22: MEASURES OF ASSOCIATION FOR NOMINAL AND ORDINAL DATA -- TAKE YOUR PICK

Norusis, Ch.10: "Crosstabulation ...CROSSTABS," only 121-127

Sanford Labovitz, "The Assignment of Numbers to Rank Order Categories," *American Sociological Review*, 35 (1970), 151-524. (On Reserve)

The world of nominal-level measures of association is quite unorderly. Some popular measures are based on chi-square, a statistic that involves probability notions and thus one that is not taken up until after the midterm. We will skip the Chi-Square based measures in this chapter of Norusis and will concentrate on the "Proportional Reduction in Error" approach of lambda.

Most statistics texts advise using ordinal measures of association for ordinal data. You instructed CROSSTABS to generate some of the most common such measures: gamma, tau, and Somer's D. These are not mentioned in Kirk, but they are described briefly in Norusis. For many years, I faithfully taught these measures, which are occasionally used in the literature, but no longer. They have two shortcomings: (1) all but gamma lacks a PRE interpretation, and (2) gamma's PRE interpretation is so strained (in terms of concordant and discordant *pairs* of cases) that it is nearly useless. I have since streamlined the course by dropping these measures, expecting that they will continue to disappear from the literature. What measures should be used instead to analyze ordinal data? I suggest using the product moment correlation for data that are clearly measured on an underlying continuous variable, even if the categories are discrete. The basis for this advice lies in the practical consequences of measurement -- as described in Labovitz's early article. Labovitz argues that it is all right to use Pearson product moment correlations with ordinal data. Why? Does the value of Pearson's r conform closely enough to the various ordinal measures of association on your CROSSTABS output to support Labovitz?

Assignment: Refer to the *lambda, tau, d, gamma,* and *r* values on the CROSSTABS output that you ran for yesterday's assignment. Study the discussion in Norusis about lambda. She does not give you quite enough information to calculate lambda in general, but it will be helpful for my discussion if you have read her section beforehand.

Using interval measures of association for ordinal data is a controversial point, and serious students may want to explore this issue further by reading this article:

Binder, Arnold. "Restrictions on Statistics Imposed by Method of Measurement: Some Reality, Much Mythology," *Journal of Criminal Justice*, 12 (1984), 467-481. (You can skip the matho on pages 473-474 and still follow the article's argument.

Binder notes that textbooks often admonish against treating ordinal data as interval in statistical analysis, but he argues that this position is misguided. Binder says that the meaning of numbers used to measure social cocnepts rests in the empirical relationships that exist among them.

> The critical point is the establishment of stable, predictable relationships among variables measured in this way makes the variables no less useful than if they were based on known scale properties. ... It would be most pleasant if the writers of introductory texts in research methods could change their treatment of measurement scale/statistics relationships. (p. 478)

Binder cites other useful references in te debate. See also:

Tufte, Edward R. "Improving Data Analysis in Political Science," *World Politics*, 21 (July, 1969), 641-654.

VI. STATISTICAL INFERENCE: ELEMENTARY PROBABILITY AND SAMPLING THEORY

OCTOBER 23: A PRIORI AND EMPIRICAL PROBABILITIES

Kirk, Ch.7: "Probability," 233-258

Statistics courses in mathematics departments devote great attention to probability theory. Unfortunately, they often pay little attention to measuring variables and measuring relationships between variables, which we emphasize. Some knowledge of probability theory, however, is crucial to statistics, especially to statistical *inference*, which enables us to make predictions to populations based on data from samples of the population. This chapter will require close reading. Be sure to do the exercises.

Assignment: Exercises in Kirk, Ch.7: 3, 5, 6, 7, 9, 13, 14, 23, 26, 27, 29, 31

OCTOBER 25: DISCRETE AND CONTINUOUS PROBABILITY DISTRIBUTIONS

Kirk, Ch.8: "Random Variables and Probability Distributions," *only* 261-278

Kirk, Ch.9: "Normal Distribution . . . Sampling Distribution," *only* 297-309

(Ch.9 uses z-scores that we discussed earlier. You may need refreshing on the standard score transformation and the normal distribution. If so, reread pp. 286-291.)

These chapters distinguish between the distribution of a variable in the *population* and the distribution as observed in a *sample*. That's easy enough to grasp, but it also discusses the (hypothetical) *sampling distribution* of means observed in an infinite number of samples. If you are a normal human being, you will find this notion of a "sampling distribution" difficult to understand at first, but you will grow to believe in it. Be *sure* you understand that the *standard error of the mean* is the standard deviation of the *sampling distribution* of the mean.

Assignment: At the beginning of the course, you ran FREQUENCIES on your original sample of ten cases that you saved as MYDATA to compute the precentages of the states' vote cast for George Bush in 1988. **BRING THIS OUTPUT TO CLASS ON THIS DAY.** Because each of you had a sample of 10 cases from the universe of states, we will be able to generate a *sampling distribution* from your individual sample means as you call them out and as I record them on the blackboard.

Assignment: Exercises in Kirk, Ch.8: 3 (a), 11 (a-e), 21, 27

Ch.9: 22, 26, 27

According to my former colleague, Philip A. Schrodt, "A billion monkeys typing a billion years on a billion typewriters will not, in fact, produce the works of Shakespeare with any reasonable probability." *See* Schrodt, "The Statistical Characteristics of Events Data," Paper prepared for the 1988 Meeting of the International Studies Association, St. Louis, page 19. Here is his explanatory footnote:

7. Assume each monkey produces, on average, one character every three seconds on a 24-hour shift. This will yield about 10^7 characters per monkey per year, so 10^9 monkeys over 10^9 years will type 10^{25} characters (25=9+9+7). A typewriter produces about 10^2 characters and if we conservatively assume all of these are equally likely (i.e. ignore the simultaneity effects of the shift key and shift lock...) the probability of randomly producing any given sequence of k characters is 10^{-2k}. The probability of never producing the target sequence in n sequences is therefore $(1-10^{-2k})^n$.

Hamlet's "Alas, poor Yorick.." soliloquy is about 1000 characters in length. At the expense of a little imprecision, we have about 10^{25} 1000-character sequences generated by the monkeys, so the probability of never finding the Yorick sequence is

$$p = (1-10^{-2000})^{10^{25}}$$

We can get a rough idea what this is equal to by noting

$$\ln(p) = 10^{25} \ln(1-10^{-2000})$$

and since $d \ln(x)/dt|_{x=1}=1$, $\ln(1-10^{-2000}) \approx -10^{-2000}$, so

$$\ln(p) \approx -10^{25} 10^{-2000} \approx -10^{-1975}$$

$$p \approx \exp(-10^{-1975})$$

which is a number which is very, very close to 1.0, even allowing for all of the approximations...

Isn't mathematics wonderful?

VII. TESTING HYPOTHESES AND PREDICTING TO POPULATIONS

OCTOBER 28:
DIFFERENCE OF MEANS TEST: SINGLE SAMPLE

Kirk, Ch.10: "Statistical Inference: One Sample," *only* 311-325
Kirk, Ch.11: "Statistical Inference: Other One-Sample Test Statistics," *only* 341-346

The difference of means test for a single sample applies when one wishes to determine if one sample (e.g., Republicans in Illinois) differs significantly from a given population (e.g., all Republicans in the nation). For the first time, you will be reading about "levels of significance" and "significant" statistics. Be sure you understand the relationship of the "alpha value" to the level of significance. Kirk begins Ch.11 by describing it as "Theme with Variations," he uses it as the vehicle to discuss the importance of testing for the population mean, μ, when the variance of the population, σ^2, is unknown. Frankly, I think that this is a confusing way to discuss the topic of sampling error of the mean, so I am putting together what he had cut apart--which will require you to jump across chapters in reading.

Assignment: Exercises in Kirk, Ch.10: 7, 11

OCTOBER 29:
ONE-TAILED AND TWO-TAILED TESTS

Kirk, Ch.10: "Statistical Inference: One Sample," *only* 326-339
Kirk, Ch.11: "Statistical Inference: Other One-Sample Test Statistics," *only* 346-352

We sometimes say that two-tailed tests are used for *nondirectional* hypotheses, while one-tailed tests are for *directional* hypotheses. Make sure you grasp this distinction between types of hypotheses. Which type should your research paper employ?

Assignment: Exercises in Kirk, Ch.10: 20 (A), 24 (a,b)

OCTOBER 30:
SIGNIFICANCE TESTS FOR CORRELATION COEFFICIENTS

Kirk, Ch.11: "Statistical Inference: Other One-Sample Test Statistics,"*only* 359-363

One of the "variations on a theme" that social psychologist Kirk sneaks into this catch-all chapter is, for us, the rather important topic of testing for the significance of correlation coefficients. Psychologists often don't compute correlation coefficients so don't rely on these tests as much as sociologists and political scientists.

Assignment: Exercises in Kirk, Ch.11: 18 and 19

NOVEMBER 1:
2/3 EXAMINATION

This examination will cover everything since the beginning of the course.

NOVEMBER 4: EVALUATION OF 3/2 EXAMINATION AND DISCUSSION OF RESEARCH PAPERS

The 2/3 examination will be evaluated according to conventional criteria of *reliability* and *validity* . I have provided a specific guide to preparing the paper for this course, and I will describe some data sets available for you to access and analyze. Although it is possible for you to analyze data that you collect on your own, I advise working with one of the data sets made available to you as SPSS files. That will enable you to concentrate on data analysis rather than on collecting data yourself, which can be enormously time-consuming.

NOVEMBER 5: DIFFERENCE OF MEANS TEST: TWO SAMPLES

Kirk, Ch. 12: "Statistical Inference: Two Samples," 371-406

More interesting research examples arise for two sample tests than for single sample tests. For example, one might test for significant differences between a sample of Democrats and a sample of Republicans, between a sample of developed nations and a sample of underdeveloped nations.

Assignment: Do exercise 4 in Kirk, Ch. 12.

NOVEMBER 6: THE T-TEST IN RESEARCH

Kirk, Ch.13: "Statistical Inference: Other Two-Sample Test Statistics," 409-413
Norusis, Ch.12: "Testing Hypotheses about Differences in Means: Procedure T-TEST," 150-165

The brief reading in Kirk has one purpose: the introduce the notion of testing for the *equality of variances* between two population with the F-test. Few research hypotheses assert such differences in variance, but the F-test figures into the T-TEST procedure in SPSS. So read Kirk to get the general idea of using the F-test, which should help you understand the computer printout.

Assignment: Using the POLITY file, prepare a directional hypothesis concerning the differences between any two regional grouping of nations on any of the interval variables in the file as your dependent variable. (You will need to use the GETINFO command to learn the variables in the POLITY data.) Because REGION is a polychotomous variable, you will have to choose only two groups or otherwise recode the nations into two groups to test for the difference between two means. By default, the T-Test program assumes that the groups are coded 1 and 2, so recode the nations into two groups before running T-TEST. (Re-read Norusis, Ch.3, on recoding if necessary.) Should you choose independent samples or paired samples? Interpret your results. You'll see some computer printout from the T-TEST program on the final exam.

Do exercises in Norusis, Ch.12: Syntax 1, 2 and Statistical Concepts 1-5, 7, 8.

VIII. ANALYSIS OF VARIANCE

NOVEMBER 8:
ONE-WAY ANALYSIS OF VARIANCE

Kirk, Ch. 15: "Introduction to the Analysis of Variance," *only* 449-471

Analysis of variance is suitable for analyzing the effects of a nominal-level independent variable (such as regions of the country) on an interval-level dependent variable (such as percent vote for the Democratic candidate for president). It is especially useful in analyzing the results of psychological experiments, in which subjects are assigned to various "treatment groups" (the independent variable), and the effects of these treatments are assessed on some variable such as memory retention, aggression, problem-solving, and so on. Analysis of variance is less frequently used in social and political research, but the ideas underlying analysis of variance are conceptually important to multiple regression analysis, which *is* a standard technique in both. In reading this chapter, be sure you understand the **within, between,** and **total** ***sums of squares.***

Assignment: Do exercises in Kirk, Ch.15: 1 and 2

NOVEMBER 11:
ONE-WAY ANALYSIS OF VARIANCE IN RESEARCH

Norusis, Ch.11: "Describing Subpopulation Differences: MEANS," 139-149
(You need know only the TABLES and STATISTICS subcommands.)
Ch.15: "One-Way Analysis of Variance: ONEWAY," 199-201 and 204-211
(You do not need know any of the subcommands.)
Kirk, Ch. 15: "Introduction to the Analysis of Variance," 494-496
Virginia P. Lacy, "Political Knowledge of College Activist Groups: SDS, YAF, and YD," *Journal of Politics*, 33 (August, 1971), 840-845. (On reserve)

Compare the ANOVA summary table on page 467 in Kirk to that in Norusis on page 201. You will have to interpret SPSS ANOVA output on the final, so better learn how to read the ANOVA table. Also understand the nature of the F-distribution, which you should be able to distinguish from the t-distribution, to which it is related. (How?) The Lacy article is an especially clear usage of analysis of variance to analyze differences among activist students.

Assignment: Run ONEWAY and MEANS for the POLITY file, using in both runs the same variable that you used previously as the dependent variable in your T-TEST assignment, but this time use REGION itself as the independent variable. Compare the results to see what analysis of variance tells you compared with the t-test, and vice versa. On MEANS, specify the ANOVA option on the STATISTICS subcommand. Then conduct the same analysis using the ONEWAY program for analysis of variance. Compare the MEANS and ONEWAY results. Are there any differences in the analysis of variance?

Do exercises in Norusis, Ch.11: Syntax 3 and Statistical Concepts 2

NOVEMBER 12:
ETA-SQUARED
AND
ANOVA
ASSUMPTIONS

Kirk, Ch. 15: "Introduction to Analysis of Variance," 471-476

Once again compare the MEANS and ONE-WAY results for your analysis variance. The MEANS procedure generates an ETA and an ETA-SQUARED statistic for analysis of variance, but the ONE-WAY procedure does not. The ETA-SQUARED statistics is a PRE statistic that--similar to R-SQUARED--states the proportion of the variance in the nations" scores on your dependent variable is "explained" by REGION.

Assignment: Compute ETA-SQUARED from the ANOVA table for the ONE-WAY results. There will be a question like this on the final exam.

NOVEMBER 13:
INTERVAL
ESTIMATION

Kirk, Ch.14: "Interval Estimation," *only* 425-431 (top of page)

I have found it hard to get the topic of interval estimation into the flow of the course, but we must do it before discussing some statistics in multiple regression, taken up next. I don't particularly care for Kirk's discussion of this topic, and I will supplement it with other material.

IX. MULTIPLE REGRESSION

NOVEMBER 15:
MULTIPLE
REGRESSION

Norusis, Ch.18: "Multiple Linear Regression Analysis: REGRESSION"
"Linear Regression," *only* 242-262 (much of this is review)
Kirk, "Ch.6: ". . . Multiple Regression and Multiple Correlation," 218-224

Multiple regression is one of the most powerful and popular techniques of multivariate analysis in social and political science. It finds the best *linear* and *additive* combination of independent variables for predicting to a single dependent variable according to the "least squares" principle of best fit. The technique assumes that the data are interval in character, but it is frequently applied to ordinal data and even nominal variables when they are reduced to dichotomies (values of 0 and 1) and treated as "dummy" variables. There are some problems associated with these departures from the classical measurement assumptions, but sensible departures seem worth the risks, for the technique is so revealing in its analysis. We will concentrate mainly on the capabilities of regression analysis, rather than on its limitations, which are treated in courses devoted to multivariate analysis. One such course is Sociology D01, which is usually offered in the winter quarter. Like a typical statistics text in psychology, Kirk's book slights multiple regression, and we will have to rely heavily on the chapter in Norusis. In reading Norusis, pay special attention to the distinction between unstandardized coefficients (B's) and standardized regression coefficients (betas). Which type is produced when the input data is in the form of z-scores?

Assignment: Do exercises in Kirk, Ch.6: 19 and 20.

NOVEMBER 18:
MORE ON
MULTIPLE
REGRESSION

Research Progress Reports are due
Norusis, Ch.18: "Multiple Linear Regression Analysis: REGRESSION"
"Multiple Regression," 262-296

We continue our discussion of multiple regression, moving to the point where you can do a multiple regression analysis of your own.

Assignment: Using the STATES file, choose five variables that you think would provide the best explanation of the states' vote for Bush in 1988. Use theory (your head) to pick the first five variables and use the computer to winnow the five down to three (or fewer). This assignment will test your understanding of the SPSS regression procedure. Follow your REGRESSION command with the subcommands WIDTH=80 and DESCRIPTIVE, and use STEPWISE. The WIDTH subcommand will format your printout for 80 columns and DESCRIPTIVE will give you MEAN, STDDEV, and CORR by default. Then specify your variables on the VARIABLES subcommand.

Do exercises in Norusis, Ch.18: Syntax 1, 2 and Statistical Concepts 1, 3, 4, 5, 10, and 11.

NOVEMBER 19:
MULTIPLE
REGRESSION
IN RESEARCH

Edward R. Tufte, *Political Control of the Economy* (Princeton University Press, 1978), pp.105-115 (On Reserve)
Lewis-Beck, Michael S. and Tom W. Rice, "Congressional Election Forecasting," *The Political Science Teacher*, (Summber, 1988), 14-16.

Edward Tufte once proposed a formula to predict the loss of seats in "off-year" elections. Several scholars have gotten into the business since then. The article by Lewis-Beck and Rice pretty much bring us up to date on the industry.

NOVEMBER 20:
MORE REGRESSION ANALYSIS AND DUMMY VARIABLES

Alan Wells, "The Coup d'Etat in Theory and Practice: Independent Black Africa in the 1960s," *American Journal of Sociology*, 79 (1974), 871-887.

Robert W. Jackman, "The Predictability of Coups d'Etat: A Model with African Data," *American Political Science Review*, 72 (December, 1978), 1262-1275. (Both articles are on reserve.)

The Wells article uses a dummy variable in a minor way. I will give a more general discussion in class. Jackman's article is much more sophisticated that anything we have read to now. Compare his analysis to that by Wells on the same topic. Which is more convincing and why? You will not understand everything in Jackman, but you should at least be familiar with most of the statistical methodology. (For those of you who are especially interested in this topic and who wish to explore it further, see the exchange of views, "Explaining African Coups d'Etat," *American Political Science Review*, 80 (March, 1986), 225-249.) If you can decipher the Wells and Jackman articles now (and couldn't before the course began), you have come a *very* long way in a few short weeks. Congratulations, you should not be able to understand 75 percent of the quantitative articles that you are likely to encounter in mainstream political science and sociology journals. If you want to understand more, take a more advanced in statistical analysis offered in political science or elsehwere in the social sciences. (D01 in Sociology is a good candidate.).

Assignment: Work on your research papers.

NOVEMBER 22:
COMPARING REGRESSION WITH ANALYSIS OF VARIANCE

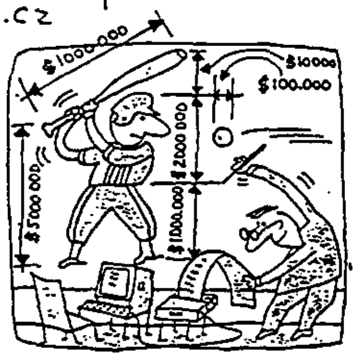
Having seen that dummy variables can be used in regression analysis, you may have suspected that there is a more general relationship between regression analysis and analysis of variance. There is, and I will demonstrate their similarities in class.

Assignment: Here is a special challenge: Using IF statements in SPSS, create "dummy" variables (those which assume a value of either 0 or 1) for 3 of the 4 regions in the American states file. Use these dummy variables to predict Bush's 1988 vote by each of the regions. This amounts to using a qualitative variable (region) in regression analysis. Then use MEANS (or ONEWAY) to analyze Reagan's vote by the single variable, region. Compare the two results for similarities and differences.

Economic Scene | Leonard Silk

NYT 6/21/91 p.C2

Predicting the Pay Of Ballplayers



As professional sport becomes big business, it is getting increasing attention from economists, rookies and otherwise. Baseball is a fertile field for economic analysis because of the comprehensive record keeping; the passionate interest of owners, players and fans in the data; the intense bargaining between players and their union and club owners, and the hot competition for talent.

*Prof. Lawrence R. Klein, a Nobel laureate in eco*nomic science at the University of Pennsylvania, found that in his last class in econometrics before his retirement this month, three of his students wanted to write term papers on sports — two on baseball, one on basketball. The best of three fine papers, in Mr. Klein's judgment, was the one turned in by Joshua A. Engel, a junior from Omaha. Professor Klein liked it because "it shows what an industrious and bright undergraduate can do these days — armed with large data banks, powerful computers (PC variety), and abundant software; it could never have been done, as a term paper, in our days."

• • •

The Engel paper's aim is to explain and predict the relationship between a ballplayer's performance and pay. It goes beyond what is regarded as the *classic study by Gerald Scully, "Pay and Per*formance in Major League Baseball," published in the American Economic Review in 1974. For batters, Mr. Scully used the lifetime slugging average (total bases hit divided by times at bat); for pitchers he used the lifetime strikeout-to-walk ratio to estimate how much marginal (extra) revenue each player contributed to his team's earnings.

But the Scully model has not been too successful in predicting players' salaries. To develop a better

the team's or the player's proposal. But there are myriad measures of performance. The issue is which are most significant in determining salaries.

After testing a host of variables, Mr. Engel found that the best *determinants of this year's salary for* hitters was a combination of last year's salary, last year's batting average, home runs hit and "runs created" — the sum of runs scored plus runs batted in minus home runs. Fielding averages and *stolen bases did not correlate with pay. For pitch*ers, he found, the best salary predictors were last year's salary, last year's earned-run average, games won, games saved, total innings pitched and total games played.

The Engel models explain about 94 percent of the salary one year ahead for major league ballplayers who filed for salary arbitration. And, after checking, it held for other players not involved in arbitration. The remaining 6 percent appears due to differences in bargaining skills or merely random error.

Since there are high transaction costs in setting baseball pay annually, clubs like multiyear contracts and players who sign them can expect to get

stra, the Phillies' centerfielder, at $2,534,000; he actually got $2,550,000.

But there were some big misses. Mr. Engel told a classmate that on the basis of his performance criteria, Roger Clemens, the Boston Red Sox pitching ace, was really worth $10 million a year. But he signed a four-year contract worth $5,380,250 a season. Fans gasped at the price, but Red Sox owners said they had got a bargain. In 1990, he had an e.r.a. *of 1.93 (2 runs less than the average league pitcher* — a phenomenal record). In baseball, productivity clearly pays; the contribution to a team's success means everything, and without their star pitcher, the Boston Red Sox would not be a serious pennant contender.

Mr. Engel found that when all other variables were held constant, every additional home run was worth $9,000 in next year's salary, and every extra *run created was worth $6,000, for hitters. For pitch*ers, one more victory is worth $38,000; one save, $16,000; one more inning pitched, $3,000 and every one-tenth of a point off the e.r.a. is worth $12,000 on next year's pay.

• • •

In the world of business, there appears to be little or no correlation between the performance of chief *executives and their pay. Business Week reports* that in 1990, UAL's chairman, Stephen M. Wolf, collected $18,301,000 in salary, bonuses and stock-based incentive plans "for heading up a company whose profits slid 71 percent." Chrysler's Lee Iacocca got a 25 percent increase in pay last year as the company's earnings dropped 79 percent. And the average chief executive's salary and bonus climbed 8.5 percent as profits fell by 7 percent.

The big difference between baseball and most business corporations seems to be that the market for ballplayers is highly competitive and the contribution of individual players to team success a critical bargaining factor, while the pay of chief executives is commonly controlled by the executives themselves, with well-paid boards of directors dependent on the chief executive willing to go along, and individual performance records are often overlooked or explained away.

X. INTRODUCTION TO NON-PARAMETRIC TESTS

NOVEMBER 25:
NONPARAMETRIC
TESTS:
CHI-SQUARE

Kirk, Ch.17: "Statistical Inference for Frequency Data," *only* 531-548

Chi-square is one of the most common tests of statistical significance in the social science literature. Its popularity lies in its applicability to nominal data and in its intuitive basis of understanding as differences between "expected" and "observed" frequencies. As a soft-summer night insires songwriters to compose romantic ballads, chi-square inspires statisticians to devise measures of association based on chi-square's sensitivity to differences between observed and expected frequencies. Alas, songwriters have produced better results. All the various chi-square based measures of association have important shortcomings, which explains why lambda is usually the preferred measure of association for nominal variables.

Assignment: Study tables on pp. 536 and 542 in Kirk; learn how to compute chi-square. There is no substitute for computing chi-square by hand to understand thestatistic. Assume that you encounter this tableshowing the distribution of political party preferences in a ward in Evanston.

	DEMOCRAT	REPUBLICAN	INDEPENDENT
Ward 1	50	70	30
Ward 2	30	60	20
Ward 3	60	65	15

Compute the chi-square test for independence to see whether this distribution is signifcant at the .05 level. Compute the value for chi-square using this format:

	Observed	Expected			
Cell Entries:	O	E	O - E	$(O - E)^2$	$(O - E)^2/E$
Row 1, Col 1	50				
Row 1, Col 2	70				
Row 1, Col 3	30				
Row 2, Col 1	*				
Row 2, Col 2					
Row 2, Col 3					
Row 3, Col 1					
Row 3, Col 2					
Row 3, Col 3					

Chi-square = sum of the column =

*(etc., you take over and complete)

You should obtain a chi-square value of 9.70. Determine the degess of freedom and decide whether to reject the null hypothesis of independence.

NOVEMBER 26-27: **HELP ON PAPERS**	Classes these two days will be devoted to answering questions on papers.

REVIEW FOR FINAL EXAMINATION

Classes during CAS reading week will be held, but they will be devoted to reviewing the course material in preparation for the final examination. Your papers are due on **December 3.**

DECEMBER 2:	*THEORY, MEASUREMENT, AND UNIVARIATE STATISTICS*
DECEMBER 3:	*STATISTICAL INFERENCE AND HYPOTHESIS TESTING* -- **papers due**
DECEMBER 4:	*BIVARIATE DISTRIBUTIONS: STRENGTH, FORM, AND SIGNIFICANCE*
DECEMBER 6:	*MULTIVARIATE ANALYSIS*
DECEMBER 9:	**Final Examination at 9:00 am**