Sociology 541: Analysis of Sociological Data I
Spring 1999

John L. Martin

Description: The purpose of this course of to familiarize students with the range of standard statistical techniques used in sociological analysis, to allow them both to understand their range of applicability, and to critically assess their use. We will examine both the mathematical underpinnings of these techniques, and actual applications, but we spend a fair amount of time in the beginning working on the fundamental underlying concepts.

Requirements: Students will, in addition to careful reading of the textbook, be required to run weekly analyses that employ the methods; these should be written up, and may, at the student's discretion, be handed in for comments. One set of analyses, those bearing on the same set of data, will be combined into a final paper, in which the student should demonstrate the commonalities and differences in the results reached by different methods, and choose which are telling us the most about the data. Students may choose a different substantive application to do in lieu of this set.

Students familiar with basic arithmetic operations should be able to successfully complete this course; however, familiarity with algebraic notation for unknowns is helpful. A special session will be conducted for any students who need a brush up on algebra.

Readings: We use the aptly named textbook, Statistics, by Freedman, Pisani, and Purves, 3rd edition. This is a fantastic book, but it contains a lot of information. But believe it or not, not always enough, so we'll supplement it with a few additional readings from a book by Alan Agresti and Barbara Finlay, which is a bit denser. I'll put that book in the lounge; if you ever finish the Freedman, Pisani and Purves (FPP), and feel ready to learn a bit more, read the corresponding chapters in Agresti & Finlay (AF). Finally, I am not following the order of FPP. (Indeed, I just discovered that my order corresponds a bit more closely to that of Blalock's Social Statistics, which I can also make available.) This creates some difficulties, I understand, but I think the overall effect with be positive. **PLEASE BEGIN THE READINGS BEFORE THE FIRST DAY OF CLASS. WE WILL START RIGHT IN.**

OUTLINE

WEEK 1 (January 20): DISCRETE VARIABLES: TEST OF A PROPORTION

Reading: Part I (only 28 pages), and chapter 3. That is, just start reading up to page 50. We won't actually be doing much about the histogram, but it got messy to try to skip around. You are always encouraged to look at the review exercises and see if you can do them; if not, be prepared to ask questions.

We begin by asking what a variable is. We then discuss the difference between qualitative and quantitative measures, and between descriptive, inferential, and analytic statistics. Taking the case of discrete variables, we start with the simplest descriptive statistics, such as percentage. We

extend to the cross-tabulation. The difficulties enter when we use a sample from a population, leading to questions of inference. We begin experimenting with a coin-tossing type experiment, and derive a practical binomial distribution to answer an inferential question.

WEEK 2 (January 27): PROBABILITY AND INDEPENDENCE
Reading: Chaps 13, 14 in FPP; Wonnacott & Wonnacott's, Selection on Bayes's theorem, if I can find it in time.

We define probability in a common sense fashion, noting that this definition produces a frequentist interpretation in large-scale cases. We then turn to Joint probabilities, and the rule of the multiplication of probability, and the special case of independent probabilities. We conclude by discussing Bayes's theorem (though we don't stress the "theorem" part), and how to trace probability "path" models in either direction. We note that conditional probability is related to the idea of "controlling for one event".

WEEK 3 (February 3): DISTRIBUTIONS AND TEST OF PROPORTIONS
Reading: Chapter 15, FPP
We now briefly derive the mathematics behind the test of significance which we carried out in practice last week (we only use the binomial distribution, and don't introduce the test of significance involving the standard deviation). We introduce the ideas of inference, null hypothesis, statistical significance, and distribution: our test of statistical significance tests our inference from the sample to the population that the null hypothesis is true, given the distribution produced under that null hypothesis. We then also introduce the idea of confidence interval, but this is de-emphasized in favor of the notion of a critical test.

WEEK 4 (February 10): TESTS OF DIFFERENT DISTRIBUTIONS
Reading: FPP, Chapter 28, 29
NOTE: These chapters makes references to the SD and SE, which we haven't examined yet. Don't worry. They aren't central, and we'll get to them next week.
We begin with cross-classified data of the 2 by 2 table, and extend it to the R by C case. We build on weeks I and 2's terminology to understand a chi-square test. While we do not go into the asymptotic theory, we bring out the following points: (1) Just as when we were able to derive the distribution of a statistic which could be used for an inferential test of a null hypothesis, so in this case, someone has derived a statistic for the null hypothesis of independence; (2) the null hypothesis is equivalent to a model of independence, where (by last week's reasoning) the real values should be computed to the "marginal" probabilities; (3) the marginals are considered as real and fixed for the purposes of the test.

WEEK 5 (February 17): CONTINUOUS VARIABLES: THE MEAN, VARIANCE AND STANDARD DEVIATION
Reading: Chapter 4, 7 (a few pages on plots…not a big deal—we'll pick this up in a few weeks. Just read it before week 9).
The mean is accessible to most students on an intuitive level, but we work through an example in any case. We then turn to the variance and standard deviation. We have seen the utility of a distribution for a test statistic, or a distribution of observations themselves under a null

hypothesis. We now examine the idea of a distribution of an observed variable itself. Like Moliere's character who found he had been speaking prose, we find that we have been using a "multinomial" distribution for nominal data; we now move to continuous variables measured at the interval level (we make a note of the difference, and stress that these methods do not necessarily apply to non-continuous and bounded variables, even if they are interval). We define the standard deviation and the variance of a variable without justifying their need; that waits until next week.

WEEK 6 (February 24): THE NORMAL DISTRIBUTION
Reading: Chapter 5, 6 (on error)
We recall our derivation of a binomial distribution; as we examine the normal curve, we see the similarity of form. The normal curve, we learn, is important mainly because the distribution is summarized only by its mean and its standard deviation. This allows us to make a test using a confidence interval as done before, for any statistic that can be transformed into a normal deviate. It also simplifies further work if our variables have normal distributions. We then go on to tests of various null hypotheses regarding a mean in a population given sample data with a given mean and standard deviation, especially Student's t and the Z test.

WEEK 7 (March 3): THE STANDARD ERROR
Reading: FPP:Skim 16 (try to figure out what they mean by Box Model), Read 17, skim 18.
Now we get to the idea of a probability distribution, and the related idea of the standard error. Does this clarify anything from before? I don't know.

WEEK 8 (March 10): TEST OF DIFFERENT MEANS AND DIFFERENT PROPORTIONS USING THE STANDARD DEVIATION

Reading: FPP Skim Ch. 26, Read Ch. 27,
Week 6's second exercise leads to a discussion of the standard deviation of dichotomies. We then extend last week's approach to testing to cover the cases of two proportions. We notice that we have constructed a new type of statistic, which seems intermediate between the distribution of a test statistic such as a chi-square, and the distribution of a variable. This is the distribution of a difference between two variables. We extend the case to where our test is of two different means. We see that the t statistic can describe the difference between sample means, as well as between a sample mean and a null hypothesis. We review and consolidate the first half of the course, re-emphasizing the tools needed to understand regression and correlation, these being (1) the difference between the distribution of an inferential statistic and a variable; (2) the variance of a variable as describing its distribution (assuming that it is normal); (3) use of inferential tests and null hypotheses.

WEEK 9 (March 24): THE IDEA OF COVARIANCE
Reading: TO BE ANNOUNCED
We attempt to reach correlational analysis without going via ANOVA, and so we need to define the idea of covariance. We re-examine our definition of variance, and, using a scatterplot as reference, make the extension of covariation on intuitive grounds,

WEEK 10 (March 31): DEFINITION OF A CORRELATION

Reading: Chapters 8, 9

Last week's exercise demonstrates that we need to put the covariation in the context of the variance of each variable. When we do this, we have the correlation coefficient. We then turn to the "significance" of the correlation coefficient, i.e. the test of the null hypothesis that it is zero.

WEEK 11 (April 7): THE IDEA OF BIVARIATE REGRESSION; ALGEBRAIC NOTATION

Reading: Review Chap 7, sec 5, Read CH 10, Skim 12, sections 1 and 3.

The correlation, as we have seen, is a symmetric measure. But many of our arguments are causal and therefore asymmetric. We introduce the idea of an independent variable, and a dependent variable, and how to link them with an equation. We note that this implies absolute determinism, and so we discuss how to add "error" on to this equation. We also discuss the meaning of the constant term.

WEEK 12 (April 14): REGRESSION AS VARIANCE AND COVARIANCE

Reading: Same as last week, add Ch 11, sections 1-2,

We now work through the regression equation in more detail, and focus on the slope parameter, pointing to the differences and commonalities with correlation. We get an idea of what "least squares" means, and hence how a regression line is fit to data. We focus on the assumptions that are required to derive the "model" of the regression data; we are moving towards analytic statistics (or parameters), which have meaning only within a model. Finally, we note the relation between the R^2 statistic and the correlation.

WEEK 13 (April 21): INFERENCE ABOUT REGRESSION

Reading: Agresti 323-342

We are not only interested in a slope's absolute value, but whether it is significantly different from zero. We apply the technique used in Week 6 to test the null hypothesis that a slope is 0 with a Z test. We also apply the idea of confidence intervals.

WEEK 14 (April 28): THE IDEA OF MULTIPLE REGRESSION

Reading: Agresti Chapter 10.

We go through the idea of controlling for more than one independent variable at a time, and how extending the model to take into account further independent variables requires other assumptions. We discuss these non-technically. We finally discuss the multiple R^2 statistic as a goodness-of-fit measure (loosely understood).