

**INTRODUCTION TO STATISTICS  
FOR POLITICAL SCIENCE:  
1. Introduction**

Stephen Ansolabehere  
Department of Political Science  
Massachusetts Institute of Technology

Fall, 2003

# 1. Introduction

Statistical tools are essential for social scientists. Basic concepts of statistics, especially randomness and averaging, provide the foundations for measuring concepts, designing studies, estimating quantities of interest, and testing theories and conjectures.

People are not always good statisticians. It is hard to maintain discipline in observing the world. We often learn from what is convenient - a violation of random sampling. We often do not calculate averages well.

To learn these concepts with the depth associated with a graduate education - where you will have the facility to use these concepts in your own research and possibly contribute to the development of statistical models that others may use - requires some mathematics. We will use, repeatedly, three sorts of functions - polynomials (especially quadratics), exponentials, and logarithms. We will also use summation, as that is necessary for the calculation of averages, and summation comes in two forms - discrete and continuous (or integration). We will use differencing and differentiation (the continuous version of differencing). Finally, we will use probability a special branch of mathematics designed to study uncertainty.

This course is designed to be a self-contained introduction not only to the concepts but also to the tools of statistics for social sciences. At the beginning of this course I will review the basic mathematical tools used in statistics. As a result we will study calculus and probability theory. Much of the basic mathematics that social scientists use in statistical analyses and in formal modeling comes from the calculus, especially limits, derivatives, and integrals. Probability provides a theory of uncertainty, and is thus the essential tool of statistics.

## 1. Two core ideas in statistics.

### A. Averaging

Statistics involves studying the frequencies of events and behaviors.

We assume that every event has its own likelihood of occurrence, such as the likelihood

of the birth of a boy or girl. The long - run average is a measure of that frequency. One important law of statistics is the Law of Large Numbers. If we observe the repetition of a certain trial or experiment or event, such as birth, the long-run frequency with which one outcome or another happens, such as a boy or a girl is born, is extremely close to the true frequency of that outcome. A second important law of statistics is the Central Limit Theorem, which states that the frequency of possible outcomes of a sum of variables follows a bell-shaped (or normal) curve. We will make both of these laws more precise later in the course.

## B. Randomness

Probability is the study of randomness and chance. The systematic study of probability emerged as an important mathematical subject of study in the 18th Century. In the late 18th and 19th Centuries the application of probability spread beyond games of chance to the study of physical and social behavior. And in the 20th Century researchers realized that one could use randomness to increase the efficiency with which we learn. That is perhaps the most surprising and counter intuitive aspect of statistics – randomness is useful.

Two core applications of this idea are (1) random sample surveys and (2) randomized experiments.

### 1. Random Sample Surveys: How can we learn about 100 million people with just 1000?

Random sample surveys are the most widely used tool for measuring quantities of interest in all of the social sciences. Nearly all government data are collected using random sample surveys - including measures of the economic and social conditions of the nation, ranging from crime to inflation to income and poverty to public health. Random sample surveys are staples of political organizations and academics interested in understanding national opinion about important public policies and public officials.

How do random sample surveys work? A relatively small group of people are chosen at random and interviewed. The average answer to a particular question in a random sample is

taken to represent or measure the average answer to that question in the entire population from which the sample is taken. How many people are to be interviewed and what they are to be asked is a matter of choice for the social scientist. But, the power of the random sample survey is that random choice of individuals gives the researcher leverage - allowing for great economy in the study of populations.

## 2. Randomized Experiments

People have conducted controlled experiments for centuries, especially using physical objects. Controls involving creating conditions in which all other factors are held constant. Even with the best controlled experiments, it is possible to leave some potentially important factor uncontrolled. Such uncontrolled factors might create spurious relations or mask important effects. Perhaps the most profound contribution of probability theory to scientific study of social and physical behavior is the notion that random assignment of individuals to different experimental conditions (such as receiving a drug or receiving a placebo) can reduce or even eliminate the threat of spurious effects.

## 2. Fundamentals of Research Methods

### A. Measurement and Estimation

1. Concepts and Variables – the constructs or behavior we wish to understand. A good example is “inequality.” Exercise: Define inequality.

2. Measures – the mathematical representation of the concept. For example, the income distribution in a society might be used to measure inequality. Exercise: devise a measure of the total amount of income inequality in a country.

3. Measurement Theory – what requirements do we impose on our measurement device. (i) accuracy (with enough observations we would arrive at the correct answer), (ii) precision (low noise), (iii) reliability (can replicate).

## B. Model Building

### 1. Effects and Behavioral Relationships

Social scientists frequently want to measure the effect of one factor on another. There are many such examples. What is the effect of police on crime? What is the effect of additional military force on the probability of winning a battle or war? How does class size affect educational performance? How do electoral rules, such as single member districts, translate votes into legislative seats?

In each case, there is one factor whose levels or values we would like to vary, such as the number of police, in order to observe changes in a second factor, such as the crime rate. The first factor we call an *independent* variable, and the second factor, a *dependent variable*.

### 2. Accounting

We seek to make a complete accounting of behavior. In this respect we value models in which a set of variables has high explanatory power. We also demand parsimony: simpler is better.

*Example.* Housing sales prices can be predicted very well as a function of list prices. In a normal market sales prices are 92 percent of list prices, and the fit is extremely good.

### 3. Equilibrium Concepts

Many ideas and conjectures about how social relations produce outcomes: maximizing behavior, dynamic adjustment, efficient markets, or natural selection. The forces that create social outcomes make it difficult to give *causal* interpretations to observed effects or relationships.

## C. Inference

A fundamental methodological problem is knowing when you should go with one argument or idea or a competing argument or idea. When we measure phenomena, we often then

use the measurements to draw inferences about different ideas. Are data consistent with an argument or idea? What conclusions can we draw about theories from data? In the end, then, statistics involves a bit of decision theory.

Predictions of a theory or conjectures about the world are called hypotheses. When specifying hypotheses it is important to be clear about all of the possible values. In a court of criminal law, hypotheses are questions of guilt or innocence. In medicine, hypotheses are about the condition of the patient, such as whether a cancer is benign or malignant or whether a woman is pregnant or not. In the scientific method generally, the question is whether an conjecture is true or not.

Unfortunatly, we never observe the truth. We use data to make decisions about hypotheses. The evidence brought to a trial are data. A series of test are data. An academic study generates data. The problem of inference is how to use data to make decisions about hypotheses. Ultimately, that will depend on the value we place on different sorts of outcomes from our decisions. However, we can formulate the problem we face quite simply.

We want to make the correct decision, and we can make a correct decision one of two ways. First, we may decide, using the data, that the hypothesis is true and the state of the world is such that it is true. Second, we may decide, using the data, that the hypothesis is not true and the state of the world is such that the hypothesis is not true. We may also make errors two ways. We may decide that the hypothesis is true when it is in fact false or we may decided that the hypothesis is false when it is infact true.

One a central objective of researchers is to avoid either of the two sorts of errors. Statistical design is fundamentally about how to minimize the chances of making a mistaken judgment.