

POL 602

Prof. Matthew Lebo

Week 1 – August 30, 2005

Introduction to Statistics

Discussion of the Syllabus.

Statistics is about learning to quantify uncertainty. Allows us to speak with complete confidence about things we are uncertain about.

Do we know what percentage of the American public approves of the job George W. Bush is doing as president?

No, but we are 95% certain that it is between 43 and 47% according to the latest Gallup poll.

We can use this ability to learn about things that have already occurred and then use it to make predictions about things that haven't occurred.

There are three related disciplines for us as statisticians.

1. Data Analysis – gathering, displaying and summarizing data.
2. Probability – learning about the laws of chance in and out of the Casino.
3. Statistical Inference – the science of using specific data and probability theory to make general conclusions.

We'll work our way through these three as the semester goes on.

A note first on terminology. An important distinction in statistics and, more broadly, theoretical modeling, is between the *stochastic* and the *deterministic*.

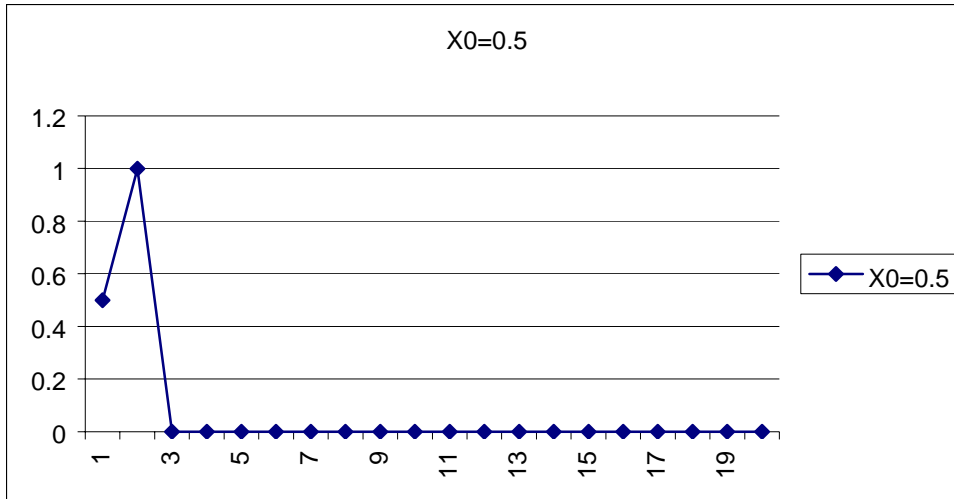
Stochastic refers to those elements or that component of a system that is random.

Deterministic refers to that component which is non-random, or *determined* by events or characteristics evident to the observer.

As an example, take a simple-looking function that could be iterated over time (a “time series” in the methods jargon—note this function is the logistic map):

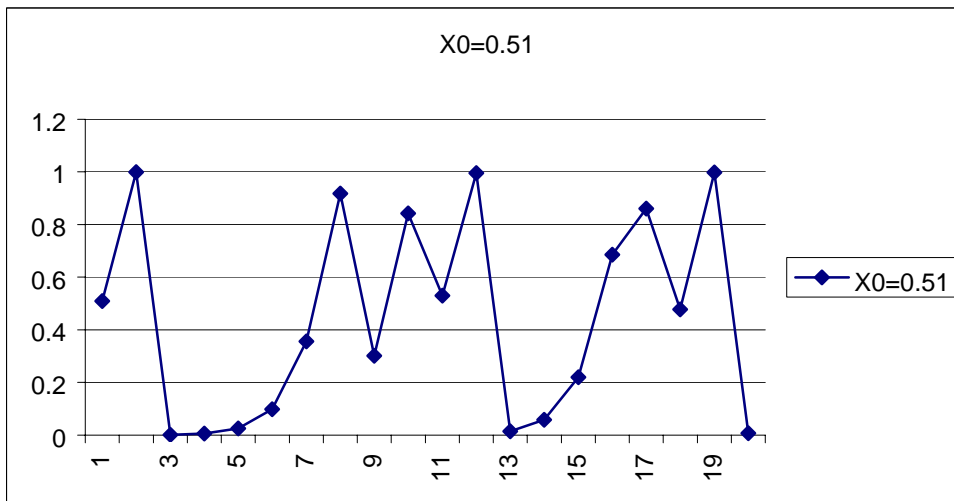
$$x_{n+1} = 4x_n(1 - x_n)$$

What does the behavior of this function look like for a given starting value? Let $x_0 = 0.5$:



Simple enough: the function goes to 1 and then drop to zero forever.

What if we choose a starting value very close to 0.5, such as 0.51?



What's going on here? There are no dice or other processes that we think of as random here?

The point is that this system is a completely deterministic one: given the starting value, we can compute where it will be exactly in any number of iterations.

Yet, to an outside observer, it might appear stochastic—random fluctuations, or “noise” as opposed to “signal.” How we think a real social system is operating, what parts of it are deterministic at our level of observation, and what parts of it are stochastic, is a central issue in your methodological training.

In this course, we focus on how to model the stochastic parts, and how to use these models to do inference—test hypotheses about social phenomena and try to reduce our ignorance about the world.

Regarding more specific statistical terminology, the first chapter in the text introduces a great many terms, some of which we shall briefly discuss today.

To some extent this is putting the cart before the horse—we will not have the mathematical foundation to really understand these terms for many weeks.

Nevertheless, they are familiar terms that most of us use in conversation (maybe) and read in the paper. Just keep in mind that we are treating them informally for now.

First is the idea of a *population*—the “universe” of possible data (real or conceptual); the people or things about which we are making inferences.

Second is the idea of a *sample*—a subset of the population. Sampling is done often due to cost and feasibility (and, occasionally given the real world, accuracy—recall the sampling vs. enumeration debate in the 2000 U.S. Census).

A third concern in statistical inference is the *degree of goodness* (as the text puts it). How well does our sample-based inference approach the true population values.

So then, to restate, the objective of statistics is to make inference about a population based on information in a sample (or many) and provide an associated measure of goodness for the inference.

This isn't easy.

If our sample isn't random we will introduce bias into our inferences.

There are famous stories of survey errors such as the 1936 presidential election.

Some more basic terminology:

Statistic: sample data characteristic (e.g. \bar{Y} , the sample mean)

Parameter: a population characteristic (Greek letters used, e.g. μ for the mean)

Sample size: n ; *Population size*: N .

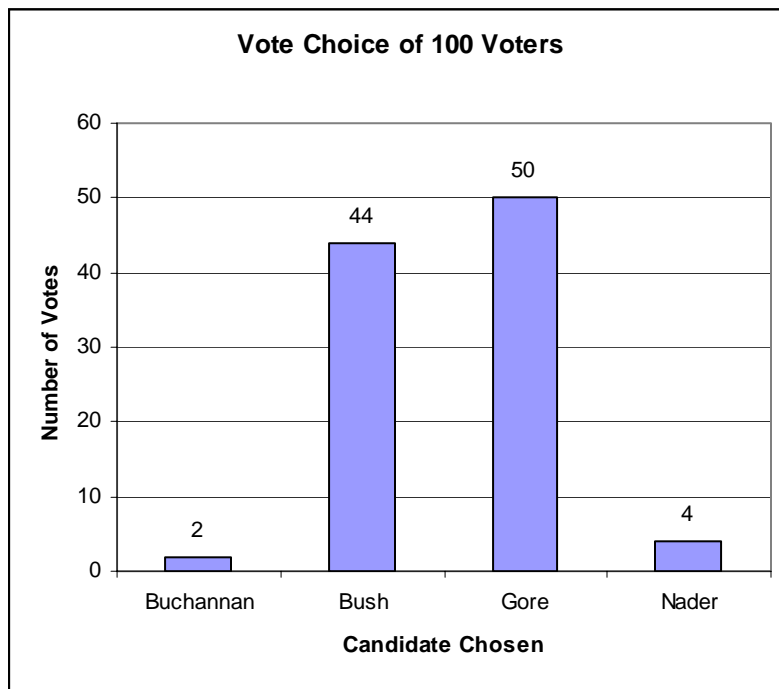
Dependent Variable: The particular quantity we are interested in. In the following model, for example, the dep.var. is vote choice:

$$\text{Vote Choice}_i = f(\text{age}_i, \text{ideology}_i, \text{gender}_i)$$

Age, Ideology, and Gender are *independent variables* (or *covariates*)

A large amount of data will seem to vary randomly (in part) as a function of these independent variables. A primary goal of inference for us is to estimate the effects or contributions of independent variables on the dependent variable.

The dependent variables, as a random variable, can be said to have a *distribution* (a great deal more one this later), which can be illustrated graphically. For a simple distribution of vote choice, we may be able to use a “bar chart,” or relative frequency histogram:



This is an example of a variable with 4 non-ordered choices (a nominal variable).

But we can have many types of variables and this frequency distribution can take on many forms.

We'll see normal distributions and many other types of distributions.

Knowing many distributions and the math behind each one of them is an important part of this course.

Why? Because different political phenomenon are measured in different ways.

The normal distribution is one that we are familiar with – we can find it in many places.

What kind of data will lead to a normal distribution?

Any value is possible – no reason to expect more above the mean than below the mean.

But when we think of actual things we wish to measure, its hard to find examples.

We may want to look at whether a judge votes for or against something – two choices gives us a binomial distribution.

Many other things we measure will have a strange DGP (data generating process) for which we'll need to know how to describe and deal with the distribution.

Much of this class you'll be asking, "Why do I need to learn another one of these?"

One answer is that the messier the data are the better the statistician needs to be and in political science, our data are very messy.

Getting our Variables Straight

Here's another example of the research process:

An article in Newsday reports the results of a study comparing 300 students from disadvantaged homes, half of whom were randomly enrolled in a pre-school program; the other half was not enrolled in a pre-school program. A follow up study twenty years later shows that 80 % of youngsters who were enrolled in the pre-school program finished high school, while only 60 % of those not enrolled in the program managed to finish high school. The results of this study are being used to argue before Congress that pre-school programs for disadvantaged children should be funded.

1. What is the dependent variable?
2. Identify the independent variable.
3. What is the hypotheses being tested?

Pre-School Enrollment Increases the likelihood of High School Graduation.

4. What about the type of relationship – is this a causal relationship?

These are important basic questions that you need to be very good at very quickly to understand your readings in other classes.

Another example: One hundred patients are told to carefully record for a month the number of calories they consume per day and the number of hours of exercise they particpate in. They are weighed at the beginning and at the end of the month.

Descriptive Statistics

Ways to characterize data.

There are two important classes of descriptive statistics that we will discuss: *measures of central tendency* and *measures of dispersion, or variation*.

First, central tendency:

The *mean* (or *arithmetic mean*) is the familiar average from elementary school mathematics. In the case of the sample, it is given by the formula:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i$$

(and the formula for μ , the population mean is the same except that N is the size).

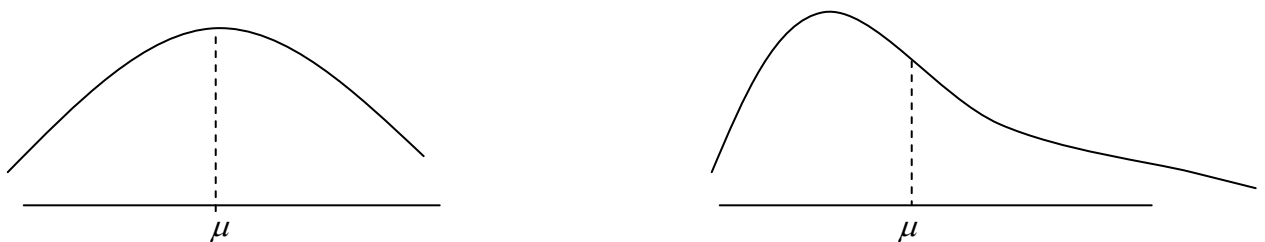
The *median* is the middle value of a distribution (equal number of observations on either side/take the average of the two middle numbers if an even number of observations).

The median can be especially important as a way to characterize central tendency if a distribution is highly *skewed* (more later) [Note: distribution of income in the U.S. example].

In 1984 the University of Virginia announced that its Department of Rhetoric and Communications graduates' *mean* starting salary was \$55,000. The reason? Outlier Ralph Sampson drafted to be a center in the NBA was included. The median wasn't published.

The final measure of central tendency we will discuss is the *mode* or *modal value*. This is simply the most commonly occurring value (especially useful for categorical data for which numerical assignments are arbitrary).

Why are we concerned about variation or dispersion? Consider the following two distributions, each with the same mean:



Clearly, more information than the mean is required for us to have a clear picture of the behavior of a random variable.

The simplest measure of variation is the *range*—the largest value minus the smallest. This is easy to calculate, but rather uninformative. E.g.: the data 5,20,20,20,20 and 5,5,5,20,20 each have the same ranges, but are clearly significantly different in distribution.

A second measure of dispersion is the *variance*—the sum of squared deviations from the mean divided by N for populations or n-1 for samples (Why? It will take until chapter 8 of the text for us to understand why this sample correction factor exists; for now note that as n and N approach infinity, (n-1) is not much different from N). More formally:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^N (y_i - \bar{Y})^2$$

A closely related measure of variation is the *standard deviation*, which is simply the square root of the variance in either the population or the sample.

The book introduces what is called the “empirical rule:” many data in the world (including the social sciences) are distributed in a bell-shape (approximately Gaussian or Normal). If this is true for a particular group of data, then we can say that:

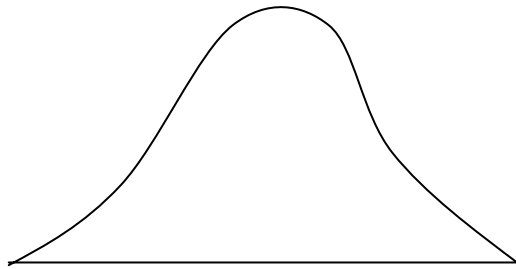
$$\mu \pm \sigma \approx 68\% \text{ of all measurements}$$

$$\mu \pm 2\sigma \approx 95\%$$

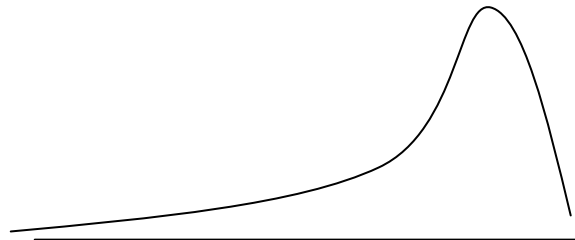
$$\mu \pm 3\sigma \approx 99.7\%$$

One final note on terminology: *skew*. We will define it more satisfactorily later, but now suffice it to say that a skewed distribution is one that is somehow asymmetrical:

Symmetrical



Positively-skewed (“skewed to the right”)



Negatively-skewed (“to the left”)

