INTRODUCTION TO STATISTICS FOR POLITICAL SCIENCE: 3. Random Variables

Stephen Ansolabehere
Department of Political Science
Massachusetts Institute of Technology

Fall, 2003

3. Probability

A. Introduction

Any good data analysis begins with a question. Not all questions are amenable to statistical analysis (though that is usually a deficiency of the question, not of statistics). Examples of interesting hypotheses abound in political science. Here are some classic and recent hypotheses that have been subjected to data analysis:

Duverger's Law "A majority vote on one ballot is conducive to a two-party system" and "proportional representation is conducive to a multiparty system." (Duverger, *Party Politics and Pressure Groups* (1972), p. 23.)

The Democratic Peace "Democracies almost never engage each other in full-scale war, and rarely clash with each other in militarized interstate disputes short of war." (Maoz, "The Controversy over the Democratic Peace," *International Security* (Summer, 1997), p. 162.)

Abortion and Crime "Children born after abortion legalization may on average have lower subsequent rates of criminality" because fewer unwanted births occur. (Donohue and Levitt, "The Impact of Legalization of Abortion on Crime," *Quarterly Journal of Economics*, May 2001, p. 380.)

These hypotheses, while controversial, are fairly typical of the type of questions that statistics might be helpful in answering.

A *testable hypothesis* is an empirical claim that can, in principle, be disconfirmed by the occurrence of contradictory data. For example, Duverger's claim that single ballot majority systems have two major parties could be disconfirmed if we found single ballot majority systems with three or four parties. There are, of course, many claims in the social sciences that are not presented in the form of testable hypotheses. Some are just vacuous, but in other

cases one just needs to do the necessary work to determine what evidence would support and contradict the hypothesis. This is not a problem of statistics and we will have little to say about it. Nonetheless, it is always useful to ask what evidence would be sufficient to contradict one's favored hypothesis. If you can't think of any, then it's probably not worth gathering the data that would *support* your hypothesis.

In practice, disconfirming a hypothesis is not nearly so straightforward as the preceding discussion would suggest. When hypotheses are of the form "If A, then B," the simultaneous occurrence of A and not B disconfirms the hypothesis. However, few hypotheses in the social sciences are of this type. Duverger, for instance, did not claim that *all* single ballot majority systems have two parties; rather, he claimed that there was a *tendency* for such systems to have two parties. It would not disprove Duverger's law for a single country to have, say, three major parties for a limited period of time. Both the U.S. and U.K. have had significant third (and sometimes fourth) parties during certain periods, but we still consider them essentially two party systems.

In fact, nearly all social science hypotheses are phrased as "If A, then B is more likely to occur." Social science hypotheses are rarely deterministic. The occurrence of A only makes B more likely, not inevitable. This is true in part because many factors other than A can influence whether B occurs. For example, workers with more schooling tend to have higher earnings than those with less, but there are many other factors, such as ability and experience, that also affect earnings. Even if we were to assemble a very long list of factors and we were able to measure each factor perfectly—two very big "ifs"—it is doubtful that we could account for all variations in earnings. For this, as for most social phenomena, there remain elements of arbitrariness and luck that we will never be able to explain.

Even when we can establish that there are systematic differences in the value of the dependent variable for different values of the independent variable, we do not know what caused these differences.

In an experiment, the analyst manipulates one variable (the so-called "independent variable") and observes what happens to the response (the "dependent variable"). If the manip-

ulations are properly designed, then one can be confident that observed effects are caused by the independent variable. For example, in medical experiments, subjects are recruited and randomly assigned to treatment and control groups. The treatment group is given a drug and the control group receives a placebo. The study is conducted *double blind*—neither the subjects nor the experimenter knows who has received what until after the responses have been recorded. The differences between the treatment and control group cannot be attributed to anything other than the treatment.

In the social sciences, controlled experiments are rare, though increasing in frequency. The values of the variables are not the result of an intervention by the analyst and the data may have been collected for a completely different purpose. Electoral systems and abortion laws are not chosen by political scientists–we only observe what happens under different electoral rules or abortion laws. We can never be sure that the consequences we observe are the result of what we think are the causes or by some other factor. In an experiment, by way of contrast, randomization of treatment ensures that treatment and control group are unlikely to differ in any systematic way.

In recent years, experiments have been conducted on a wide range of social phenomena, such as labor market policies, advertising, and voter mobilization. Economists have also used laboratory experiments to create artificial markets. Today no one contests the usefulness of experiments in the social sciences, though many topics of interest (including the three hypotheses described above) remain beyond the range of experiments.

There is a lively debate in the social sciences about natural experiments (sometimes called *quasi-experiments*). Donohue and Levitt's analysis of the relationship between abortion and crime utilizes a natural experiment: five states legalized abortion prior to the Supreme Court's decision in *Roe vs. Wade*. They compare the differences in subsequent crime rates between these five states and the remainder. However, because the decision to legalize abortion was not randomized, many other factors might be responsible for differences in crime rates between these types of states.

Whether the data come from an experiment or an observational study, the *method of*

comparison is used to determine the effects of one variable on another. For simplicity, suppose the independent variable is dichotomous (*e.g.*, the type of electoral system is one ballot simple majority or proportional representation; the political system is democratic or authoritarian; a state either legalized abortion before *Roe vs. Wade* or not) and the dependent variable is quantitative (*e.g.*, the proportion of vote received by the top two parties; the number of wars between a pair of countries; the murder rate per 100,000 population). The method of comparison involves comparing the value of the dependent variable under the two conditions determined by the independent variable. If these values differ, we have evidence that independent variable affects the dependent variable.

Because of the inherent randomness of social processes, the method of comparison cannot be applied directly. Data are "noisy," so we must distinguish between systematic and random variation in our data. The fundamental idea of probability is that random variation can be eliminated by *averaging* data. One or two observations can be discrepant, but averaging a large number of independent observations will eliminate random variation. Instead of comparing individual observations, we compare averages.

Even if we can establish that there are systematic differences in the dependent variable when the independent variable takes different values, we cannot conclude that the independent variable *causes* the dependent variable (unless the data come from a controlled experiment). In observational studies, one cannot rule out the presence of *confounding variables*. A confounding variable is a third variable which is a cause of both the independent and dependent variable. For example, authoritarian countries tend to be poorer than democracies. In this case, level of development could be a confounding variable if richer countries were more willing to fight poorer countries.

Statistics provides a very useful technique for understanding the effects of confounding variables. Ideally, we would be able to run a controlled experiment where, because of randomization, the independent and confounding variable are unrelated. In observational studies, we can use *statistical controls* by examining the relationship between the independent and dependent variable for observations that have the same value of the confounding variable.

In the Democratic peace example, we could examine the relationship between regime type and war behavior for countries at the same level of development.

B. Types of Data and Variables

In the examples described above, we have encountered various different types of data. Roughly data can be classified as being either *qualitative* or *quantitative*. Qualitative data comes from classifications, while quantitative data arises from some form of measurement. Classifications such as gender (male or female) or party (Democratic, Republican, Green, Libertarian, etc.) arise from categorizations. At the other extreme are physical measurements such as time, length, area, and volume which are naturally represented by numbers. Economic measurements, such as prices and quantities, are also of this type.

We can certainly attach numbers to categories of qualitative variables, but the numbers serve only as labels for the categories. The category numbers are often referred to as "codes." For example, if Democrats are coded as 1, Republicans as 2, Greens as 3, and so forth, you can't add a Democrat and a Republican to get a Green. Quantitative measurements, on the other hand, are inherently numerical and can be manipulated like numbers. If I have \$10 and you have \$20, then together we have \$10 + \$20 = \$30.

In between purely qualitative variables, such as gender, and obviously quantitative variables, such as age, are rankings and counts. For example, ideology might have three categories (liberal, moderate, and conservative), which we might score as 1, 2 and 3, or as 0, 3 and 10, or any other three numbers which preserve the same ordering. In this case, the numbers are not entirely arbitrary since they represent an ordering among the categories. Similarly, we might count the occurrences of a qualitative variable (*e.g.*, whether someone attends church) to obtain a quantitative variable (*e.g.*, how many times someone attended church in the past year).

There is a large literature in properties of measures, but it has surprisingly little relevance for statistical practice. Most statistical methods for quantitative measurements can

be adapted to handle qualitative measurements without too much difficulty. Furthermore, measures that might seem naturally quantitative (such as years of schooling) may be just rankings (or worse) of the variables of theoretical interest (how much more educated is someone with 16 years of schooling than someone with 12?).

C. Probability and Randomness

Most people have an intuitive understanding of probability. We understand what it means when a meterologist says that "the probability of rain is 70%," though we probably don't understand how the probability was calculated. For statistics, it is necessary to be able to do some elementary (and occasionally not so elementary) probability calculations. In this lecture, we start by formalizing the concept of probability and deriving some of its properties. Although the mathematical ideas are very simple, they lead to a surprisingly rich theory.

[Mosteller studied the meaning of expressions such as "almost surely," "likely," "probably," "possibly," "might" and so forth. He found that a people give fairly consistent quantitative interpretations to these terms.]

Probability is the key concept in the development of statistics, as we will see in the next section (D). We will treat data as random variables – variables for which each value of the variable occurs with a specific frequency or probability. Data analysis consists of studying the frequency with which the values of a particular variable may occur.

1. Random Experiments

A *random experiment* is some process whose outcome cannot be predicted with certainty in advance. The easiest examples are coin tosses (where the outcome is either heads or tails) or the roll of a die (where the outcome is the number of points showing on the die). These examples may convey the misimpression that probability is primarily applicable to games of

chance. Here are some other examples more relevant to social science.

> Random Sampling. We have a population of N elements (usually people) and we draw a sample of n elements by sampling without replacement. If we have a listing of the population, we can index the population elements by $i = 1, \ldots, N$. To select the first element in the sample, we draw an integer at random from $\{1, \ldots, N\}$. Denote the element selected by i_1. We then draw the second element i_2 from the remaining $N - 1$ integers and continue until we have selected n elements. An outcome is the set of integers selected $\{i_1, \ldots, i_n\}$.

This example is *repeatable* in that we could use the same procedure over again to draw another sample (and, presumably, to get a different outcome). However, random experiments need not be repeatable as the following examples illustrate.

- How many missions can a space shuttle complete before a catastrophic failure occurs?
- How many earthquakes of magnitude 3.0 or higher will California have during the coming year?
- What are the chances of a student with high school GPA of 3.8 and SAT scores of 1500 being admitted to Stanford? MIT?
- Who will win in the California recall?

The first is an example of *destructive testing*. Once the failure has occurred, we cannot repeat the experiment. When the application involves something expendable (such as a light bulb), we may repeat the experiment with different objects—but this involves an implicit assumption that the other objects are similar enough to be somehow comparable.

In the second example, we might consider data from past years, but this again involves an assumption that chances of an earthquake don't vary over time. For Stanford and MIT admissions, it wouldn't make sense to repeat the experiment by having the same student reapply the following year (since different admission standards are applied to high school

applicants and college transfers). We could consider the admission outcomes of all students with the same GPA and SAT scores, but this does not constitute any kind of random sample.

The last example is a case where there is no obvious precedent (and, some hope, no repetition either!). Yet one could not deny that considerable uncertainty surrounds the recall and statistics, as the science of uncertainty, should have something to say about it. We shall, in fact, use random experiments to refer to any kind of process involving uncertainty.

2. The Sample Space

In a well-defined random experiment, we should be able to list the possible outcomes. It makes very little difference whether the number of outcomes is small or large. The critical issue is whether we are able to describe all possible outcomes and, after the experiment is completed, to determine which outcome has occurred.

Let ω denote an outcome and let W denote the set of all possible outcomes, so $\omega \in W$. The set W is called the *sample space* of the experiment. The outcomes ω are sometimes called *sample points*.

The sample space W is an abstract set. This means that the outcomes in the sample space could be nearly anything. For example, in the experiment of flipping a coin the sample space might be $W = \{Heads, Tails\}$.

For the experiment of drawing a sample of size n from a population of size N without replacement, the sample space might be the set of all n-tuples of distinct integers from $\{1, \ldots, N\}$,

$$W = [(i_1, \ldots, i_n) : i_j \neq i_k if j \neq k\}.$$

In general, it is more convenient if the outcomes are numerical. Thus, we may choose to relabel the outcomes (*e.g.*, $Heads = 1$ and $Tails = 0$).

Two people analyzing the same experiment may use different sample spaces, because their descriptions of the outcomes may differ. For example, I might prefer the outcomes from drawing a random sample without replacement to indicate the order in which the

sample elements were selected. You, on the other hand, might not care about the order of selection and let two samples with the same elements selected in different order to be considered the same outcome. There is no "correct" description of the sample space. It only matters that the sample space contain all possible outcomes and that we be able to decide which outcome has occurred in all circumstances.

An example from the study of government formation is instructive.

A parliamentary government is formed by the party that has a majority of seats in the legislature. When no party has a majority of seats, a coalition government may form. In theories of bargaining, it is conjectured that Minimum Winning Coalitions will form. A coalition is minimum winning if the coaltion is winning (has more than half of the seats) but the removal of any one party makes the coalition losing. Assume that parties cannot be divided–they vote as a bloc.

Consider the sample space for the formation of coalitions for the following situations.

- a Party A has 100 seats, Party B has 100 seats, and Party C has 1 seat.
- b Party A has 70 seats, Party B has 60 seats, Party C has 50, party D has 10.
- c Party A has 90 seats, Party B has 35 seats, Party C has 30 seats, and party D has 26 seats.
- d Party A has 90 seats, Party B has 50 seats, Party C has 36 seats, Party D has 30 seats, and Party E has 25 seats.

Answer. If we don't care about the order in which coalitions form, the unique coalitions are: (a) AB, AC, BC, (b) AB, AC, BC, (c) AB, AC, AD, BCD, (d) AB, AC, ADE, BCD. If we care about the order then we have a longer list: (a) and (b) AB, BA, AC, CA, BC, CB, (c) AB, BA, AC, CA, AD, DA, BCD, CBD, CDB, etc.

This idea is extended further by political theorists in the analysis of bargaining power within legislatures. Many situations produce the same sets of coalitions. These situations

are equivalent and their equivalence is represented by the "minimum integer voting weights." For example, the situation where A has 4 seats, B has 3 seats, and C has 2 seats produces the same sample space (set of possible coalitions) as the situation where A has 100, B has 100, and C has 1. The minimum integer voting weights are the smallest integer representation of the votes of parties in a coalition game. In the example where the possible minimum winning coalitions are AB, AC, and BC, the game with the smallest weights that yields those coalitions arises when A has 1 vote, B has 1 vote, and C has 1 vote. Hence, the minimum integer weights are 1, 1, 1. Many theoretical analyses begin with the definition of the game in terms of minimum integer voting weights (see Morrelli APSR 1999). This makes theorizing easier because bargaining leverage is a function of what coalitions one might be in, and only indirectly depends on the number of seats one has.

In situations (b), (c), and (d) the minimum integer voting weights are (b) 1, 1, 1, 0, (c) 2, 1, 1, 1, (d) 4, 3, 3, 2, 1.

3. Events

Because individual outcomes may include lots of irrelevant detail, we will be interested in agglomerations of outcomes, called *events*. For example, let the event E_i denote the event that element i is one of the population elements selected for our sample,

$$E_i = \{(i_1, \ldots, i_n) \in W : i_j = i for some j\}.$$

For instance, if the population size $N = 3$ and we draw a sample of size $n = 2$ without replacement, the sample space is

$$W = \{(1,2), (1,3), (2,1), (2,3), (3,1), (3,2)\}$$

and the event E_3 that element 3 is selected is

$$E_3 = \{(1,3), (2,3), (3,1), (3,2)\}.$$

In general, an event E is a subset of the sample space - . We will use the notation

$$E \subset W$$

to indicate that E is a subset of W. E could contain a single outcome ($E = \{\omega\}$), in which case E is sometimes called a *simple event*. When E contains more than one outcome, it is called a *composite event*.

Two events have special names in probability theory. The empty set $\emptyset$ is an event (the event with no outcomes) and is called the *impossible event*. Similarly, the entire sample space W, which contains all possible outcomes, is called the *certain event*. Since the experiment results in a single outcome, we are guaranteed that $\emptyset$ never occurs and W always occurs.

4. Combining Events

From set theory, you are familiar with the operations of intersections, unions, and complements. Each of these has a natural probabilistic interpretation which we will describe after we have reviewed the definitions of these operations.

The *intersection* of sets E and F, denoted $E \cap F$, is composed of the elements that belong to both E and F,

$$E \cap F = \{\omega \in W : \omega \in E and \omega \in F\}.$$

For example, suppose $W = \{1, 2, 3, 4, 5\}$. If $E = \{1, 2, 3\}$ and $F = \{2, 3, 4\}$, then $E \cap F = \{2, 3\}$.

The *union* of sets E and F, denoted $E \cup F$, is composed of the elements that belong to either E or F or both,

$$E \cup F = \{\omega \in W : \omega \in E or \omega \in F\}.$$

In the preceding example, $E \cup F = \{1, 2, 3, 4\}$. Note that we use "or" in the non-exclusive sense: "$\omega \in E$ or $\omega \in F$" includes outcomes where both $\omega \in E$ and $\omega \in F$.

The *complement* of E, denoted E^c, is composed of the elements that do not belong to E,

$$E^c = \{\omega \in W : \omega \notin E\}.$$

Continuing the preceding example, $E^c = \{4, 5\}$. Note that we need to know what elements are in the sample space - before we can compute the complement of a set. The complement is always taken relative to the sample space which contains all possible outcomes.

As mentioned previously, the set theoretic operations of intersection, union, and complementation have probabilistic interpretations. If E and F are events, the event $E \cap F$ corresponds to the co-occurrence of E and F. $E \cup F$ means that either E or F (or both) occurs, while E^c means that E does not occur. In the California recall example, let E denote the event that Governor Davis is recalled and F the event that Bustamante receives the most votes on the replacement ballot. Then E^c is the event that Davis is not recalled, $E^c \cup F$ means that the Democrats retain the governorship, and $E \cap F$ is the event that Bustamante becomes governor.

Two events E and F are said to be *mutually exclusive* or *disjoint* if

$$E \cap F = \emptyset.$$

It is impossible for mutually exclusive events to occur simultaneously. In other words, they are incompatible. In the California recall example, the events "Bustamante becomes governor" and "Schwarzenegger becomes governor" are mutually exclusive. One may occur or neither may occur, but it is impossible for both to occur.

5. Probability Measures

A *probability measure* assigns a number $P(E)$ to each event E, indicating its likelihood of occurrence. Probabilities must be between zero and one, with one indicating (in some sense) certainty and zero impossibility. The true meaning of probability has generated a great deal of philosophical debate, but, regardless of one's philosophical position, any probability measure must satisfy a small set of axioms:

First, like physical measurements of length, probabilities can never be negative.

Axiom I (Non-negativity). For any event $E \subset W$, $P(E) \geq 0$.

Second, a probability of one indicates certainty. This is just a normalization–we could equally well have used the convention that 100 (or some other number) indicates certainty.

Axiom II (Normalization). $P(\text{-}\ ) = 1$.

Third, if two events are mutually exclusive, then the probability that one or the other occurs is equal to the sum of their probabilities.

Axiom III (Addition Law). If E and F are mutually exclusive events, then

$$P(E \cup F) = P(E) + P(F).$$

(When - is not finite, we will need to strengthen this axiom slightly, but we bypass these technicalities for now.)

It is surprising how rich a theory these three simple (and intuitively obvious) axioms generate. We start by deriving a few simple results. The first result is that the probability that an event does not occur is equal to one minus the probability that it does occur.

Proposition For any event $E \subset \text{-}$, $P(E^c) = 1 - P(E)$.

Proof. Since $E \cap E^c = \emptyset$, E and E^c are mutually exclusive. Also, $E \cup E^c = W$. By the Addition Law (Axiom III),

$$P(E) + P(E^c) = P(E \cup E^c) = P(W) = 1,$$

so

$$P(E^c) = 1 - P(E).$$

The next result states that the impossible event has probability zero.

Proposition $P(\emptyset) = 0$.

Proof. Since $\emptyset = \text{-}^c$, the preceding proposition and Axiom II imply

$$P(\emptyset) = 1 - P(W) = 1 - 1 = 0.$$

Probability measures also possess the property of *monotonicity*.

*Proposition*If $E \subset F$, then $P(E) \leq P(F)$.

Proof. Note that $F = E \cup (E^c \cap F)$ and $E \cap (E^c \cap F) =$ so, by the Addition Law,

$$P(F) = P(E) + P\left(E^c \cap F\right) \geq P(E)$$

since $P\left(E^c \cap F\right) \geq 0$.

The following result generalizes the Addition Law to handle cases where the two events are not mutually exclusive.

Proposition $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

Proof Since $F = (E \cap F) \cup (E^c \cap F)$ where $E \cap F$ and $E^c \cap F$ are mutually exclusive,

$$P(F) = P(E \cap F) + P(E \cap F^c)$$

or

$$P(E \cap F) = P(F) - P(E \cap F).$$

Also $E \cup F = E \cup (E^c \cap F)$ where $E \cap (E^c \cap F) = \emptyset$, so $P(E \cup F) = P(E) + P(E^c \cap F) = P(E) + P(F) - P(E \cap F)$.

The Addition Law also generalizes to more than two events. A collection of events $E_1, \ldots, E_n$ is *mutually exclusive* or *pairwise disjoint* if $E_i \cap E_j = \emptyset$ when $i \neq j$.

Proposition If the events $E_1, \ldots, E_n$ are mutually exclusive, then

$$P(E_1 \cup \ldots \cup E_n) = P(E_1) + \ldots P(E_n).$$

Proof. The case $n = 2$ is just the ordinary Addition Law (Axiom III). For $n > 2$, the proof is by induction on n. Suppose the result holds for n, and

$E_1, \ldots, E_{n+1}$ are mutually exclusive. Then the two events $F = E_1 \cup \ldots E_n$ and E_{n+1} are mutually exclusive (since $E_i \cap E_{n+1} = \emptyset$ for $i = 1, \ldots, n$ implies $F \cap E_{n+1} = \emptyset$) and $P(E_1 \cup \ldots \cup E_{n+1}) = P(F \cup E_{n+1}) = P(F) + P(E_{n+1}) = P(E_1) + \ldots + P(E_n) + P(E_{n+1})$, where second equality follows from the ordinary Addition Law and the last equality follows from the induction hypothesis.

Last, we prove a property called *subadditivity*.

Proposition. $P(E \cup F) \leq P(E) + P(F)$.

Proof. From the Generalized Addition Law, $P(E \cup F) = P(E) + P(F) - P(E \cap F) \geq P(E) + P(F)$ since $P(E \cap F) \geq 0$.

Question of the Day. Prove Bonferroni's Inequality,

$$P(E \cup F) \geq P(E) + P(F) - 1.$$

D. Random Variables

In nearly every field of social science, we wish to predict and explain the frequency with which social phenomena occur. Doing so requires that we carefully define the variables of interest. What is the range of possible outcomes that can occur? How likely or frequent are specific events? Variables characterized this way are called random variables: the values of the variables occur with a frequency or probability.

Examples

- Votes. The vote for a candidate or party reflects the frequency with which people vote and the frequency with which they support a particular candidate or party. It may be the case that every individual is of a known type: voter or not voter; supporter of a given party or not. The support for a party in an election will reflect the frequency of voters and the frequency of supporters. It may also be the case that individuals are

not certain about their voting behavior and vote by chance occurence and may even choose a "random" strategy for their behavior.

- Educational Achievement. There are many ways that we measure educational achievement, including years of schooling completed Standardized tests are used to measure educational performance or achievement of each student. Students' abilities to answer questions vary depending on innate ability, preparation, and conditions of the testing. Also, there is an element of true randomness because students may guess at answers.

- Crime. In the population at a given time are people who have been victims of crimes. The FBI measures that amount of crime through the Uniform Crime Reports – a system of uniform reporting of criminal complaints filed with local police departments. This, of course, may not capture all crimes, as many go unreported. A second measure of the crime rate is the victimization rate, which consists of the responses to random sample surveys that ask individuals (anonymously) if they have been the victims of specific types of crimes.

D.1. Definition

More properly a random variable is a function that assigns a probability to each element and set in the sample space, - . The random variable X gives a numeric value and a probability to each element in - . In statistics we tyically use capital letters toward the end of the alphabet to denote random variables, i.e., X, Y, and Z. Occassionally, lowercase Greek letters are used. In regression analysis, ϵ denotes a particular random variable corresponding to the "residuals" or unexplained random component in the regression.

Three important features of random variables are the metric and the support. The *metric* of the variable is the sort of numbers assigned. Some variable are continuous, such as money, distance, or shares, and we use the metric of the real numbers to represent the units of measurement. Some variables are discrete. They may be ordered, as with ranks or counts, or not, as with categories.

The *support* of the variable X is the set of possible values that it can take. For example, if X is the number of heads in n tosses, then $S = \{0, 1, 2, \ldots, n\}$. Another example is that the random variable may take any value on the real number line. The support, then, is $[-\infty, +\infty]$. We will sharpen the definition later, but for now we will think of the support of X as being the range of X. Random variables are always real-valued, so their support is always a subset of the real numbers.

The *probability function* defines the frequency of specific values of X. The probability functions of random variables follow the laws of probability. The probability of any element in the support of X is between 0 and 1. The probability of *an* event in the support of X is 1. And, if we divide the entire support of the variable into n exclusive sets, then the sum of the probability of those n sets is 1. The *probability distribution*, sometimes used interchangeably with the probability function, maps the probability function for all values of X.

Unfortunately, two distinct notations are used for random variables that take continuous values and random variables that take discrete values. When X is discrete, we write the probability function as $P(X = k)$, where k are the specific values. When X is continuous, we write $f(x)$ as the probabilty density associated with the specific value of X, i.e., x. We will develop the concept of the probability function for each type of variable separately. Before doing so, we think it important to address a common question.

Where does randomness in random variables come from?

There are three potential sources of randomness.

- Nature.
- Uncertainty.
- Design.

First, Nature. Most of what we learn and know derives from observation of a particular behavior of interest to us. The concept of a random variable is useful for thinking about the fequency with which events occur in nature. Indeed, probability is often taken as a model of

individual behavior. Take the example of crime. Many forces affect whether an individual commits a crime; however, much about crime, especially crimes of passion, appears arbitrary or random. In a given circumstance an individual might "act as if tossing a coin." The field of game theory conjectures the use of "randomized strategies" – called mixed strategies. When individuals are in a situation where one action does not clearly dominate all others, their best approach is to take a given action with a probability chosen so as to make the opposition indifferent as to the player's choice. In this way, people might even inject randomness into social phenomena.

Second, Uncertainty. Researchers are uncertain about the adequacy of their models and measurements of nature. Randomness and random variables are one way of representing the researcher's uncertainty. What is not observed, it is hoped, behaves in "random" ways? By this we mean that it is unrelated to what is in our models of behavior. So, randomness is often a working assumption for empirical research.

Third, Design. The power of statistics comes from the use of randomness as an instrument for learning. Often it is infeasible to do an ideal study. It is much too expensive to conduct a census to ascertain public attitudes on every issue of interest. Instead, researchers conduct interviews with a relatively small subset of the population. Historically, there were two approaches to survey research: random samples and representative samples. Representative samples were drawn by instructing interviewers to find a certain number of people matching a set of demographic characteristics (e.g., 500 men and 500 women; 250 college educated persons, 500 with high school education, and 250 with less than high school education). Random sampling consists of drawing a random subset of individuals from a list of all individuals. Many surveys, for example, are conducted by randomly dialing phone numbers. By injecting randomness into the process, the researcher can *expect* to draw an appropriately representative sample – even of characteristics that may not have been thought important in representative sampling. We will establish exactly why this is the case later in the course. Random sampling has become the standard approach for survey research.

Experimentation is another important research design where randomness is essential. The

idea of a causal effect is that we observe the outcome variable when the treatment is present and when it is not present, everything else held constant. "Everything else held constant" is, of course, a very substantial caveat. Strictly speaking, everything else cannot be held constant. One cannot observe the same person both taking a treatment and not taking a treatment at any given time. The researcher can try to make the conditions under which the treatment is applied as similar as possible to the conditions under which no treatment is taken. Alternatively, one can randomly choose when the treatment is taken *and* when it is not taken. We expect that the effect of the treatment will be the same as if all else is held constant.

The application of randomness to the design of studies are two fundamental and surprising advantages of probability. We will treat these as two important examples as we proceed with the development of random variables.

D.2. Types of Variables

Probability functions assign probability to specific values of the random variable X. There are an infinite variety of probability functions, but in practice statisticians focus on just a few. We will learn the most fundamental – Bernoulli and Binomial (and Poisson), Uniform and Normal.

Probability functions are central to statistics. Statistics consists of three core activities – summarization of data, estimation, and inference. We use probability functions to represent random variables – that is, to summarize data. We will also need to estimate the features of probability functions. Probability functions depend on two sorts of numbers – values of random variables and constants. The constants are called *parameters*. Parameters are sometimes equal to quantities that we wish to use immediately, such as probabilities, means, and variances. Estimation consists of using data to make our best guess about parameters of the probability functions of interest. Finally, we will use probability functions to make inferences. Given the estimates we have made using the data at hand, how likely is a

hypothesized relationship to be correct or adequate?

Two broad classes of probability functions are discrete and continuous functions. We'll consider these here.

D.2.a. Discrete

A wide range of discrete random variables are encountered in practical research.

- Categorical Variables. Examples include groups of people, such as ethnicities or religions.
- Ordered Variables. Examples include rankings, such as school performance from best to worst.
- Interval Variables. Examples include counts of events, such as the number of accidents or wars.

To each value of the variable assign a number, indexed $X = k$, $k = 0, 1, 2, 3, ...$ We write the probability assigned to each value of X as $P(X = k)$. We require

$$\sum_{k=0}^{\infty} P(X = k) = 1$$

i. Bernoulli Random Variables

The building block of discrete random variables is the *Bernoulli Random Variable*. The variable X has two values or categories, such as voted or did not vote. We, arbitrarily, assign $k = 1$ as the value for one category and $k = 0$ the value for the other category. $P(X = 1) = p$, where p is a specific number and, since the probabalities of all events sum to one, $P(X = 0) = 1 - p$. This function can be written in a more compact form:

$$P(X = k) = p^X(1 - p)^{(1-X)}$$

Note: If $X = 1$, the function returns $P(X = 1) = p$. If $X = 0$, the function returns $P(X = 0) = (1 - p)$.

ii. Binomial Random Variables

Many sorts of data consist of sums of independent *Bernoulli Random Variables*, called Bernoulli trials. The probability function is called the *Binomial Distribution.* Each random variable may equal 0 or 1. When a Bernoulli trial equals 1 it is called a "success." Let $X_1, X_2, ...X_n$ be a series of independent random variables and the sum of these variables is $X = X_1 + X_2 + ... + X_n$. The support of this random variable is the set of integers from 0 to n.

What is the probability that $X = k$? That is, what is the probability that there may be k successes out of n trials? We can translate this into a set problem. The event of interest is k successes and $n-k$ failures. Let us choose one particular sequence of such events. Suppose that the first k attempts all led to success and the subsequent $n-k$ trials all led to failure. The probability of this specific sequence is $pppp...p(1-p)(1-p)...(1-p) = p^k(1-p)^{(n-k)}$. This, of course, is just one sequence of possible outcomes, and the series of k successes and $n-k$ failures may not have occurred in exactly this sequence. There are, in fact, many ways in which a subset of exactly k successes may occur in n trials. But, we do not care about the order in which the k successes occur. To calculate express the binomial probabilities we must account for the different ways that events might have occurred. From Pascal's triangle we know the number of subsets with k in and $n-k$ not in equals the binomial coefficient: $\binom{n}{k}$. Hence,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

We may map out the entire distribution function by calculating the probabilty function for each value of k.

Consider two numerical examples. First, a sequence of 10 trials where the probability of success is .5. The part of the Binomial distribution that depends on p equals $.5^k.5^{n-k} = .5^n = .5^{10} = .00097656$. The binomial coefficients require calculation of factorials. The numerator of $\binom{n}{k}$ is $10! = 3,628,800$. The denominator depends on the value of k. The result of $\binom{n}{k}$ is shown in the second column of the table. The probability distribution is as follows

Binomial Probability Calculations			
		$P(X=k)$	
k	$\binom{n}{k}$	$p=.5$	$p=.1$
0	1	.001	.349
1	10	.010	.387
2	45	.044	.194
3	120	.117	.057
4	210	.205	.011
5	252	.246	.001
6	210	.205	.001
7	120	.117	.000
8	45	.044	.000
9	10	.010	.000
10	1	.001	.000

You can verify that the probabilities sum to 1. The two distributions look quite different depending on the value of p. When $p = .5$ the Binomial distribution is symmetric and centered at 5, which equals np. Deviations from $p = .5$ create asymmetries or skew in the distribution. When $p = .1$, most of the probability concentrates in $k = 0$ and $k = 1$. Indeed, the modal value is $k = 1$, which equals np and values of k above 6 receive extremely small probability weight.

STATA provides you with a list of probability functions (type **help probfun**). The values of the Binomial Probability function can be generated using the command **Binomial(n, k, p)**. The numbers n and p are parameters of the distribution function corresponding to the number of trials and the Bernoulli probability and k is the value of random variable X.

APPLICATION: Random Sampling With Replacement

Suppose we draw a random sample of size n from a population of size N with replacement. Suppose further that M of the N objects in the population possess some characteristic. For example M people approve of the job that the president is doing and $N - M$ people do not. We would like to measure the support for the president in the population. A census is not worth the cost. Suppose further that we have a list of all people – say the list of all

registered voters or the list of all phone numbers – and we assign each individual a number. We randomly choose n numbers, say from a bin containing all N numbers or using a random number generator. Each time we draw a number we write down the individual and *replace* the number into the bin. (So the number might be drawn again.) We continue drawing individual numbers this way until there are n numbers. Of the n numbers in the sample there we may have anywhere from 0 to n supporters of the president (this is the support). What is the probability of getting a sample where k of the n sample members approve of the job the president is doing?

We know that there are N^n possible ordered samples that we could draw with replacement. How many of these samples have k persons who approve of the job the president is doing? In selecting the sample, we have n cells or positions to fill. How many ways can we fill k of these cells with "approvers" and $n-k$ with presidential "non-approvers."

To solve this problem we use the Fundamental Principle of Counting. Divide the task into three subtasks:

- We select k of the n cells to receive approvers. The remaining $n-k$ cells will receive non-approvers. There are $\binom{n}{k}$ ways to complete this task.
- We fill the k cells allocated to approvers with population members who support the president. Because we are sampling with replacement, there are M choices for each of the k cells or a total of M^k ways to complete this taks.
- We fill the remaining $n-k$ cells with non-approvers. By the same logic there are $(N-M)^{(n-k)}$ ways to complete this task.

This implies that there are

$$\binom{n}{k} M^k (N-M)^{(n-k)}$$

samples with k approvers and $n-k$ non-approvers when we sample with replacement.

We are now in a position to compute the probability of getting exactly k approvers in a sample of size n drawn with replacement. Since each ordered sample is equally likely, the

probability is calculated by dividing the number of favorable outcomesby the total number of outcomes in the sample space:

$$P(X=k) = \frac{\binom{n}{k} M^k (N-M)^{(n-k)}}{N^n} = \binom{n}{k} \left(\frac{M}{N}\right)^k \left(\frac{M-k}{N}\right)^{(n-k)} = \binom{n}{k} p^k (1-p)^{(n-k)}.$$

This expression summarizes the probability of choosing exactly k approvers (or sometimes we say exactly k successes) when sampling with replacement. It is assumed that we know M or p. If these are not known then we must estimate the probabilities. How might we estimate p?

This is an excellent expample of the general problem of statistical estimation. An important intellectual leap in statistics is to see how we can use probability to summarize data *if the parameters are known*, and we can then, chose the values of the unknown parameters to make the data most likely. This is Fisher's method of maximum likelihood. We will consider it now to foreshadow how data can be used to measure parameters generally.

The Binomial probability function above expresses the probability of observing exactly k successes in n independent trials given the value of the parameter p. Suppose that we have conducted a survey and observed k successes in a sample of n. How can we use this information to make our best guess about p? We will now treat p as a variable. Importantly, it is not a random variable; rather it is a number we will vary so as to find the value $\hat{p}$ that makes it most likely to observe $P(X=k)$.

To find the value of p that maximizes $P(X=k)$, calculate the first and second derivatives of $P(X=k)$ with respect to p. The first order condition for a maximum is

$$\frac{dP(X=k)}{dp} = \binom{n}{k} \left[k\hat{p}^{k-1}(1-\hat{p})^{n-k} - (n-k)\hat{p}^k(1-\hat{p})^{n-k-1}\right] = 0$$

We can simplify this by dividing by $\binom{n}{k}\hat{p}^{k-1}(1-\hat{p})^{n-k-1}$. The equation then simplifies to $k(1-\hat{p}) - (n-k)\hat{p} = 0$. This yields

$$\hat{p} = \frac{k}{n}.$$

The distribution that we derived from sampling with replacement, then, led naturally to a formula for estimating the unknown parameter.

iii. Hypergeometric Probabilities

Closely related to the Binomial Probabilities are the Hypergeometric Probabilities. These arise in a variety of settings, including the famous capture-recapture problem and sampling without replacement. I will present them here as the case of sampling without replacement.

The permutations allow us to compute the number of ordered sample when selection is made without replacement. Now the question is how many of these involve exactly k successes, e.g., have exactly k people who approve of the job the president is doing? Because we are sampling without replacement, $k \leq M$ and $n - k \leq N - M$ for this number to be positive.

As before, we proceed using the Fundamental Principle of Counting.

- Pick k cells or positions to hold the "approvers" and allocate the remaining $n-k$ cells for non-approvers. There are $\binom{n}{k}$ ways to pick these k cells.
- There are $\frac{M!}{M-k!}$ wasy to fill the k cells for approvers.
- There are $\frac{(N-M)!}{(N-M-(n-k))!}$ ways to fill the $n-k$ cells for non-approvers.

Thus, we the total number of ways we can select an ordered sample with k supporters and $n-k$ opponents is

$$\binom{n}{k}\frac{M!}{M-k!}\frac{(N-M)!}{(N-M-(n-k))!} = n!\binom{M}{k}\binom{N-M}{n-k}.$$

It follows that the probability of selecting a sample with exactly k successes is the number of ways of selecting an ordered sample with k successes and $n-k$ failures divided by the number of ways of drawing samples of size n without replacement:

$$P(X=k) = \frac{n!\binom{M}{k}\binom{N-M}{n-k}}{N!/(N-n)!} = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

These are called the *Hypergeometric Probabilities*.

Comments. Sampling without replacement and sampling with replacement lead to very different formulae for the probability of exactly k successes in n trials. The difference between these formulae stems from the fact that the Binomial probabilities result from a sequence of independent trials and the Hypergeometric probabilities result from a sequence of dependent trials, where the dependence arises because once an individual is chosen he or she is set aside and cannot be selected again.

It turns out that if the sample size is less than 10% of the population size the difference between the Binomial and Hypergeometric probabilities is of little consequence. The ratio n/N is called the sampling fraction. For public opinion surveys, it is rare for the sampling fraction to exceed 1%, much less 10 %.

Finally, except when the sample size is quite small (20 or less) the above factorials will be enormous, making it difficult to make the calculations above. Instead, we will use approximations. Importantly, using Stirlings formula leads immediately to the normal distribution as an approximation to the Binomial.

APPLICATION: Estimation and The Census Undercount

Every decade the U.S. Census conducts an enumeration of the population. These data are used to determine the apportionment of U.S. House seats, the distribution of public welfare money, and many other important government functions. The enumeration seeks to count all persons in the United States as of April 1 of the census year. It is known that the enumeration misses some people. The question is, how many?

To address this issue the census draws a sample of census tracts and during the subsequent summer attempts to enumerate all people in these randomly chosen areas. The Post Enumeration Survey (or PES) involves a much more intense attempt at enumeration, including looking for homeless people. Assume there are N people in the population, M people in

the original enumeration, and n people in the PES. What is the probability of observing k people in both? Again we divide the counting into parts: We compute the number of ways of observing k people in the PES who were in the enumeration and the number of ways of observing $n-k$ people in the PES who were not in the enumeration. We divide this by the number of ways that we could have drawn the post enumeration sample from the total population, yielding a hypergeometric probability:

$$P(X=k|N) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

The Census actually observes M, n, and k, but not N, which is an unknown parameter of the distribution. However, as with the Binomial problem, we can estimate N by finding the value of N, call it $\hat{N}$, that makes the probability function as high as possible (again, the principle of maximum likelihood). That is, find $\hat{N}$ such that $P(X=k|\hat{N}) = P(X=k|\hat{N}-1)$. To do this we solve:

$$\frac{\binom{M}{k}\binom{\hat{N}-M}{n-k}}{\binom{\hat{N}}{n}} = \frac{\binom{M}{k}\binom{\hat{N}-1-M}{n-k}}{\binom{\hat{N}-1}{n}}$$

This looks much nastier than it is. Writing out the Binomial coefficients fully, there are a large number of cancellations and simplifications. First, $\binom{M}{k}$ cancels. Second, $(n-k)!$ and $n!$ canel, leaving:

$$\frac{(\hat{N}-M)!/(\hat{N}-M-n-k)!}{\hat{N}!/(\hat{N}-n)!} = \frac{(\hat{N-1}-M)!/(\hat{N}-1-M-n-k)!}{(\hat{N}!-1)/(\hat{N}-1-n)!}.$$

This further reduces to: $(\hat{N}-M)(\hat{N}-n) = \hat{N}(\hat{N}-M-n-k)$. Solving for $\hat{N}$:

$$\hat{N} = \frac{Mn}{k}$$

This estimate has an intuitive appeal. From the PES, the Census arrives at an estimate that someone is contacted through the best efforts of the Census, which is k/n. The number of people enumerated equals the population times the probability of being detected by the Census enumerators. Hence, the population equals the number of people enumerated divided by the probability of being enumerated.

APPLICATION: Inference and Fisher's Taste Testing Experiment

R.A. Fisher demonstrated the principles of randomized experimentation and inference with the following simple example – tailored to contemporary tastes.

A friend states that she can distinguish between Pepsi and Coke. You bet that she can't and agree to friendly wager. You will set up 8 glasses. Four glasses are chosen to have Pepsi and four Coke. All eight are presented to your friend in random order (chosen with a probability mechanism). Of course she won't know which cups have which softdrink, but she will know that there will be four of each. If she is just guessing, what is the probability of identifying all of the glasses correctly?

This can be calculated as the hypergeometric probabilities. The task before our friend is to divide the glasses into two groups of 4. If she is just guessing, then she arbitrarily divides the glasses into 4 guesses of one type, leaving 4 of the other type. There are 70 ways of choosing a group of 4 objects out of 8 ($= \binom{8}{4}$). This is the number of points in the sample space. The task before our friend is to separate glasses in to Coke guesses and not Coke guesses; once she has made 4 Coke (or Pepsi) guesses the remaining guesses are determined. This consists of two tasks: the number of ways of guessing Cokes that are in fact Cokes and the number of ways of guessing Cokes that are infact Pepsis. There are $\binom{4}{k}$ ways of guessing k Coke glasses that are in fact Coke and $\binom{4}{4-k}$ ways of guessing $4-k$ Coke glasses that are in fact Pepsis. For example, suppose three Coke guesses are in fact Cokes and 1 Coke guess is in fact Pepsi. These can occur as follows: actual Cokes are guessed to be CCCP, CCPC, CPCC, and PCCC (or 4 choose 3 right); and actual Pepsis are guessed to be PPPC, PPCP, PCPP, and CPPP (or 4 choose 1 wrong).

$$P(X=k) = \frac{\binom{4}{k}\binom{4}{4-k}}{\binom{8}{4}}$$

Hence, $P(X=4) = 1/70, P(X=3) = 16/70, P(X=2) = 36/70, P(X=1) = 16/70$, and $P(X=0) = 1/70$. In other words if our friend is guessing she has a 1 in 70 (about one and a half percent) chance of getting all of the glasses right and a 16 in 70 chance (about

23 percent) of getting exactly three Coke and three Pepsi glasses right. About half the time she will get half of the glasses right.

Comments. This is the distribution of likely outcomes under an assumed hypothesis (that our friend is guessing) once we have introduced a random element into the test. We can use this distribution to make inferences and draw conclusions. If our friend guesses all right then we would be convinced of her ability. We should not be surprised at all if our friend gets two Coke and two Pepsi right–about half the time that should occur with guessing. If our friend identifies 3 Coke and 3 Pepsi right, then we might also be unconvinced. The probability that she got at least three (3 or 4) right is almost one in four if she is just guessing. The distribution of likely outcomes, then, is useful in establishing *in advance* what you agree is a fair condition or wager. As a skeptic you might request that the probability of correctly identifying a set of glasses be very small before you are convinced.

Such is the case generally with statistical inference. Before empirical researchers are convinced that there is something systematic occuring – that is, beyond what might occur just by chance or because of guessing – we expect that to have observed a very low likelihood of observing the data given some null hypothesis. This is called a level of confidence. Typically we set the level of confidence at .05. That is, if the null hypothesis were true, the probability of observing a deviation from the expected pattern is less that .05, or a one in twenty chance. This level is just a convention. You might demand a higher level of confidence, such as a one in one hundred chance.

We can build on this example further. Experimental design and research design generally involves choosing an appropriate number of observations so as to be able to distinguish the hypotheses of interest. In this very simple problem, no fewer than 6 glasses would suffice. With 6 glasses there is a one in 20 chance of identifying all glasses correctly if one is just guessing.

iv. Poisson Probabilities

A further discrete proability function with considerable practical applications is the Poisson distribution. The Poisson distribution is used to model rare event, such as suicide (Durkheim), accidents, and war (Mansfield).

Suppose that a rare event that occurs at a rate per time unit in a population, such as 4 suicides per year in a city of 100,000. We can divide the units of an arbitrarily large sample (population times time) into n intervals, such as per year in a city of size 100,000. The rate of occurence is denoted $\lambda = np$.

The probability of observing k occurence of a rare event in a given time interval is expressed as:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Poisson Probability Calculations

	$P(X=k)$		
k	$\lambda = .5$	$\lambda = 1$	$\lambda = 4$
0	.606	.368	.018
1	.303	.368	.073
2	.076	.184	.147
3	.013	.061	.195
4	.002	.015	.195
5			.156
6			.104
7			.060
8+			.052

The Poisson distribution can be derived as the limit of the Binomial as n tends to infinity where $np = \lambda$.

$$Bin(n, k, \lambda/n) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

We can express this again as,

$$Bin(n, k, \lambda/n) = \binom{n}{k} \lambda^k n^{-k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}$$

Let us consider the limit as $n \to \infty$. The expression $\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}$ and the term λ^k is constant. The expression $\left(1 - \frac{\lambda}{n}\right)^{-k} \to 1$. The remaining term is $\binom{n}{k} n^{-k}$. Expanding terms

we discover that this ratio tends to 1. Hence the limit of the Binomial where the number of divisions n becomes large is the Poisson.

Mansfield (World Politics, 1988) studies the incidence of war over the last 500 years in Europe as reported by various scholars and cumulated into three data sets. One dataset shows about .4 new wars per year. One might take this as a simple null hypothesis. Wars occur at random based on some underlying probability that people fight each other. The conceptual model is that conflict is part of the human condition, which has been relatively unchanged in Europe over the last 500 years.

To test this idea we can compare the actual incidence of wars to the pattern predicted under the hypothesis. This base rate implies that two-thirds of all years have no new wars; twenty-seven percent of all years are expected most years have one new war; five percent of the years are predicted to have two new wars; and .8 percent are predicted to have more than two new wars a year.

The data are strikingly close to this simple argument.

D.2.b. Continuous Random Variables

A second class of random variables take continuous values. Importantly, we will use continuous variables to approximate many discrete variables. The notation differs slightly.

D.2.b.1. Definition

Let X be a continuous random variable and x be a specific value. The probability function is written $f(x)$ and is called the probability density function. The support of continuous random variables is assumed to be the entire number line unless otherwise specified.

The density at any one point of X is vanishingly small, and we care not just about a single point but also values of the random variable in a neighborhood of points. Typically, then, we will deal with the cumulative or total probability in an interval or up to a specific point. That involves the integration of all values of the density function between two points, say a and b. We write the density as follows:

$$F(b) = \int_{-\infty}^{b} f(x)dx$$

The axioms and rules of probability apply to density functions as well as discrete probability functions. The density and cumulative probability always lie between 0 and 1. The probability assigned to any two disjoint intervals or points equals the sum of the two probabilities. And, the probability assigned to the entire support of X is 1. That is,

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Also, the probability assigned to the values of two independent continuous random variables equals the product of their densities: If X_1 and X_2 are independent then $f(x_1, x_2) = f(x_1)f(x_2)$.

D.2.b.2. Specific Distributions

i. Uniform

A uniformly distributed random variable has constant density for all values. This density has three purposes. First, it is handy for instruction. Second, computer programs and formal theorists often use it as an approximation to other distributions. Third, it can be used to generate random numbers when drawing random numbers for random assignments in experiments and random samples in surveys.

Let X be uniform on the interval a to b. The density function is constant, so $f(x) = c$. The constant is called a constant of integration: it's value is such that the density sums to 1. What is the value of c?

$$\int_a^b c dx = xc|_a^b = bc - ac = c(b-a) = 1$$

Hence, $c = \frac{1}{b-a}$. The uniform density is:

$$f(x) = \frac{1}{b-a}$$

if $a < X \leq b$ and 0 otherwise.

The density up to any point x is

$$F(x) = \frac{x}{b-a}$$

Example. How might we actually draw random numbers to construct random samples? Suppose we wish to draw 20 random numbers with replacement from 100. We can use a random number generator. In STATA **uniform()** returns uniform random numbers on the interval 0 to 1 for each observation in space dedicated to the variables. To generate random numbers in the interval 0 to 100, we multiply the random numbers by 100. To get integer values, we can round the numbers to the nearest integer. The following commands accomplish these tasks.

set obs 100000

gen x = uniform()

replace x = 100*x

replace x = round(x,1)

The first 20 numbers (indeed any 20 numbers) will be 20 random numbers from 0 to 100 drawn with replacement?

ii. Normal

The central distribution of statistical analysis (indeed, perhaps the only one you really need to know) is the Normal Distribution. This distribution has a bell-shaped curve defined by the following formula.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The parameters μ and σ are key to statistical modelling. These correspond to the mean and standard deviation of the distribution function.

Of particular importance is the standard normal. If we let $Z = \frac{X-\mu}{\sigma}$, then

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

The cumulative distribution function cannot be expressed in closed form and, in stead, must be approximated analytically. So long as you can "standardize" a problem, by subtracting the mean and dividing by the standard deviation, you can use the standard normal to calculate the probability of an interval or event. Your text book includes a table of standard normal deviates. In section you should review how to read the tables.

An important notation commonly used for the normal distribution is $N(\mu, \sigma)$ and for the standard normal $N(0, 1)$, because the standard normal has $\mu = 0$ and $\sigma = 1$. Another notation sometimes used for the standard normal is the Greek letter ϕ. The standard normal density is sometimes written as $\phi(z)$ and the standard normal distribution $\Phi(z)$.

Comments.

The normal is important because sums of random variables are approximated by the Normal Distribution. This means that for large enough n the Binomial and other distributions can be approximated by the Normal. To use this approximation, let $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. The table shows the approximation with $n = 20$ and $p = .5$. I present the part of the distribution over the values $4 \geq k \leq 10$. The distribution is symmetric around 10 and the probability weight to values less than 4 and greater than 16 is extremely small.

Normal Approximation to the Binomial

	$P(X = k)$, $n = 20$	
k	Binomial	Normal
4	.005	.005
5	.015	.015
6	.037	.036
7	.074	.073
8	.120	.120
9	.160	.161
10	.176	.178

Sums of random variables of unknown distribution are known to follow the normal. The Binomial approximation is a special case of the Central Limit Theorem. The process of adding and averaging regularly leads to the normal distribution. This allows researchers to build statistical models based on the normal distribution.

Normal random variables are very commonly assumed in modeling data – that is, in making abstract statements about data.

Example. The distribution of Votes.

For a very wide range of problems the normal distribution describes observed data extremely well.

In the US, the Republican and Democratic political parties together win over 90 percent of the vote in any election. Political scientists study American elections as if there were only two parties, because the two major parties completely dominate the electoral process. The

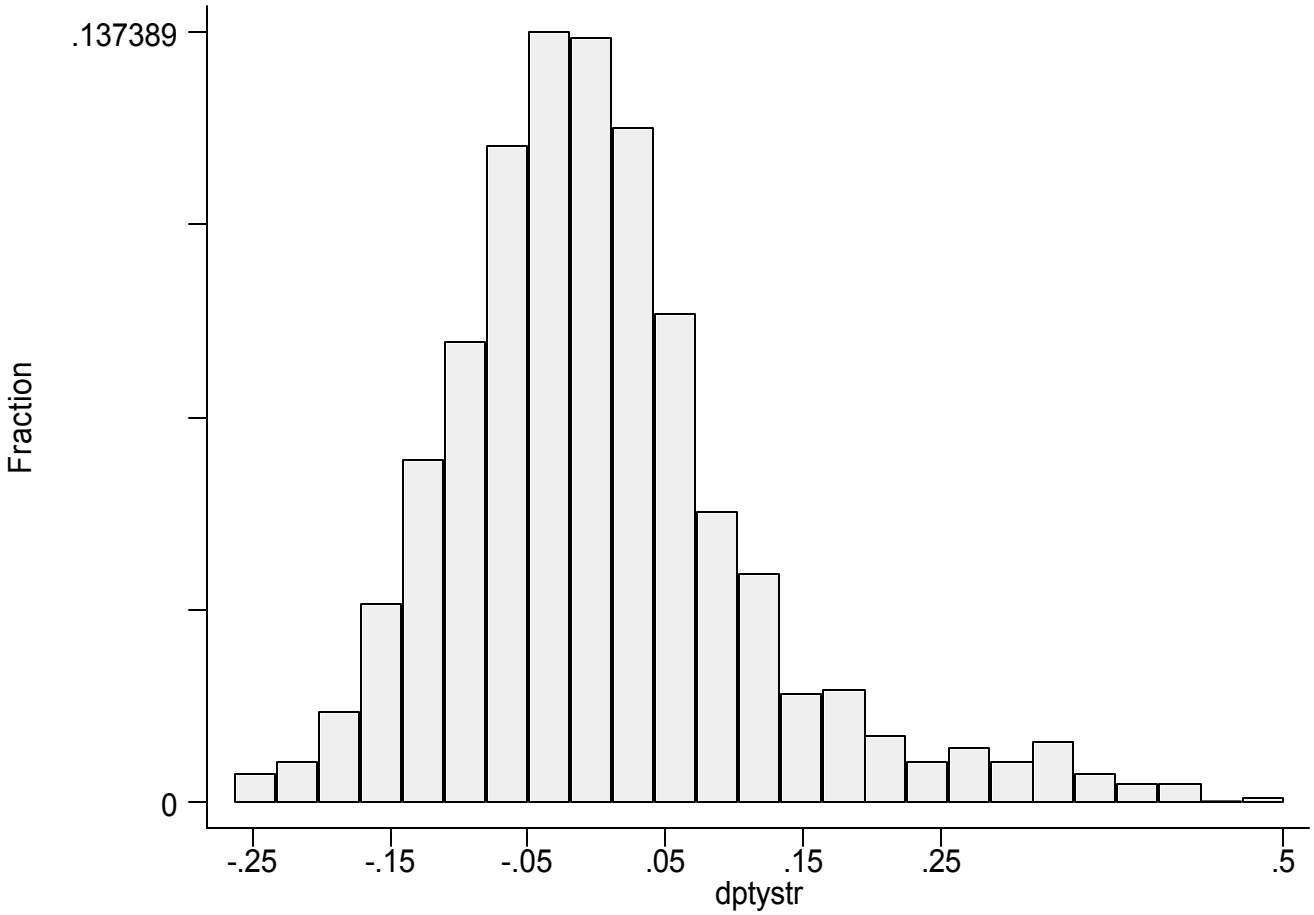
normal distribution is a very good model of the distribution of vote across electoral districts or counties in the United States. We may think of any individual voter as either a D or an R.

The vote for any congressional seat, then, is the sum of many voters. This number of D voters will follow the Binomial. Since congressional districts typically have around 300,000 voters, the number (and percent) of D voters will be approximately normal *within* districts. The distribution of votes *across* congressional districts is not so easily derived from first principles. There are a number of complications–different candidates running in different districts, variation in voters' preferences across areas, etc.

One concept is the *normal vote*. If the same two candidates ran in every district, how would their vote vary across districts? The presidential vote involves the same two candidates running in all districts. The attached graph shows the distribution of the Presidential vote across congressional districts. The data are for presidential elections from 1988 to 1996. I have subtracted the average Democratic vote in each year, so values near 0 correspond to congressional districts where the electoral division is 50-50.

The presidential vote across congressional districts follows the normal distribution extremely well. This is an important baseline for other analyses. Of note, the data deviate somewhat from normality in the upper tail. These are districts that An important theoretical question is what is the process by which the distribution of the vote across districts is normal?

Kendall and Stuart (British Journal of Sociology 1951) provide an excellent discussion of these issues.



E. Transformations

Many statistical problems involve transformations of a variable. If we know the distribution function of the untransformed variable, what is the distribution function of the resulting variable? The distribution of the transformed data will depend on the original distribution function and the transformation.

For example, many economic data are converted into a common currency. To compare income in 1960 with income in 2000, we adjust for inflation of prices. To compare income or

government accounts in one country with income or government accounts in another country we translate into the currency of one of the countries using the exchange rates. These are examples of "rescaling" data.

The general method for deriving a new distribution for monotonic transformations can be derived from the cumulative function. Suppose that $Y = h(X)$ defines the transformation and that $F(\cdot)$ is the cumulative function of X. Let $G(\cdot)$ be the distribution function of Y. We can express the density function $g(y)$ in terms of $f(\cdot)$ and $h(\cdot)$ as follows:

$$g(y) = \frac{dG}{dy} = \frac{dF(x(y))}{dx}\frac{dx}{dy} = f(h^{-1}(y))\frac{dx}{dy},$$

where $h^{-1}(y) = x$ is the inverse of the transformation evaluated at the value of Y and $x(y)$ expresses X as implicitly a function of y.

i. Linear Transformations

A linear transformation consists of multiplying a constant times a random variable and adding a constant to that product. That is, we create a new variable Y such that

$$Y = a + bX$$

What is the distribution of Y? Adding a constant to a random variable changes the *location* of the center of the distribution – shifting all values by the same amount. Multiplying by a constant changes the *scale* or distance between values. For example if the values were originally 1, 2, 3, etc., and we multiply by 2, then the new values are 2, 4, 6, etc.

Using the result above, $X = (Y - a)/b$, so

$$g(y) = \frac{1}{b}f((y - a)/b)$$

Comments. This transformation is very important to understand for using the normal distribution, because it reveals how the standard normal relates to any normal. Often we have normally distributed data and we wish to calculate the probability of observing a value bigger than a specific amount or in a particular interval. The linear transformation formula

shows how the standard normal relates to the geneal normal. Let Y be the data and Z be a standard normal random variable. Suppose that a and b are numbers such that $Y = a + bX$. Then,

$$g(y) = \frac{1}{b} f((y-a)/b) = \frac{1}{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-a}{b})^2}$$

Because the last expression is exactly the form of the general normal, we know that $b = \sigma$ and $a = \mu$ and $g(\cdot)$ is the general normal distribution function.

ii. Quadratic Transformations

The general idea for monotonic transformations can be extended to non-monotonic transformations, but we must be careful about the values of the transformed variable. The most important non-monotonic transformation in statistics is the square of a random variable.

In constructing hypothesis tests, we will usually want to know if an observed value deviates substantially from an hypothesized value. The hypothesized value is a constant and the data are summarized by a random variable. The squared distance is an obvious measure of the deviation. How do compute the probability of a big deviation?

For simplicity, let's ignore the hypothesized value. Let $Y = X^2$. How does this transformation affect the data? It does two things. (1) It folds the data. All negative values of X are now positive: $-1^2 = +1$. (2) It spreads the values greater than 1 and compresses the values less than 1: $\frac{1}{2}^2 = \frac{1}{4}$ and $2^2 = 4, 3^2 = 9, 4^2 = 16$, etc.

The folding of the data means that we need only look at the positive values of the squareroot of Y in the transformation forumla. But, all else about the formula is the same:

$$g(y) = \frac{dG}{dy} = f(\sqrt{y}) \frac{dx}{dy} = \frac{1}{2} \frac{1}{\sqrt{y}} f(\sqrt{y}),$$

for $y > 0$. If $y = 0$ the function is not defined.

The square of a standard normal random variable results in the Chi-square distribution:

$$g(y) = \frac{1}{2} \frac{1}{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sqrt{y})^2} = \frac{1}{\sqrt{y} 2\sqrt{2\pi}} e^{-\frac{1}{2}y},$$

for $y > 0$.

The normal gives rise to a family of distributions depending on the transformation of the Normal variable. The Chi-square is one such distribution and it is essential in hypothesis testing. In inference we use several such distributions. These are mentioned in the table below.

Normal Family of Distributions

Transformation	Distribution	Application
Sum of RVs	Normal	Means, Diff of Means
Square of Normal RV	χ-squared	Variance, Squared Error
Normal/ $\sqrt{\chi^2}$	T-Distribution	Mean/Standard Deviation
χ^2/χ^2	F-Distribution	Ratio of Variances Comparing Models

F. Joint and Conditional Distributions

Distributions of a single variable are extremely useful in inference and many basic problems (such as the distribution of votes or the census undercount), but most often we are interested in the relationships among several variables. What is the effect of one variable on another variable? For example, how does information conveyed through the news reports affect people's attitudes toward public affairs? How does one variable predict another? For example, can we predict election outcomes with a couple of readily observed variables, such as economic growth?

We will examine such complicated problems in depth later in the course. Here I will define the the key probability concepts and give a couple of examples.

Given two random variables X and Y, the joint distribution defines the probability weight assigned to any value described by the pair $(x,\ y)$. That is, $f(x, y)$ is the density associated with the point $X = x$ *and* $Y = y$. The cumulative is the integral over both random variables up to the level $X = x$ and up to the level $Y = y$:

$$F(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u, v)dudv$$

Also, because this is a proper probability function:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u, v)dudv = 1$$

Conditional probabilities are the probabilities assigned to values of one variable, say X, given a particular value of the other variable, say y. We write this as $f(x|Y = y)$ or just $f(x|y)$.

Case 1. Discrete Random Variables: Tables.

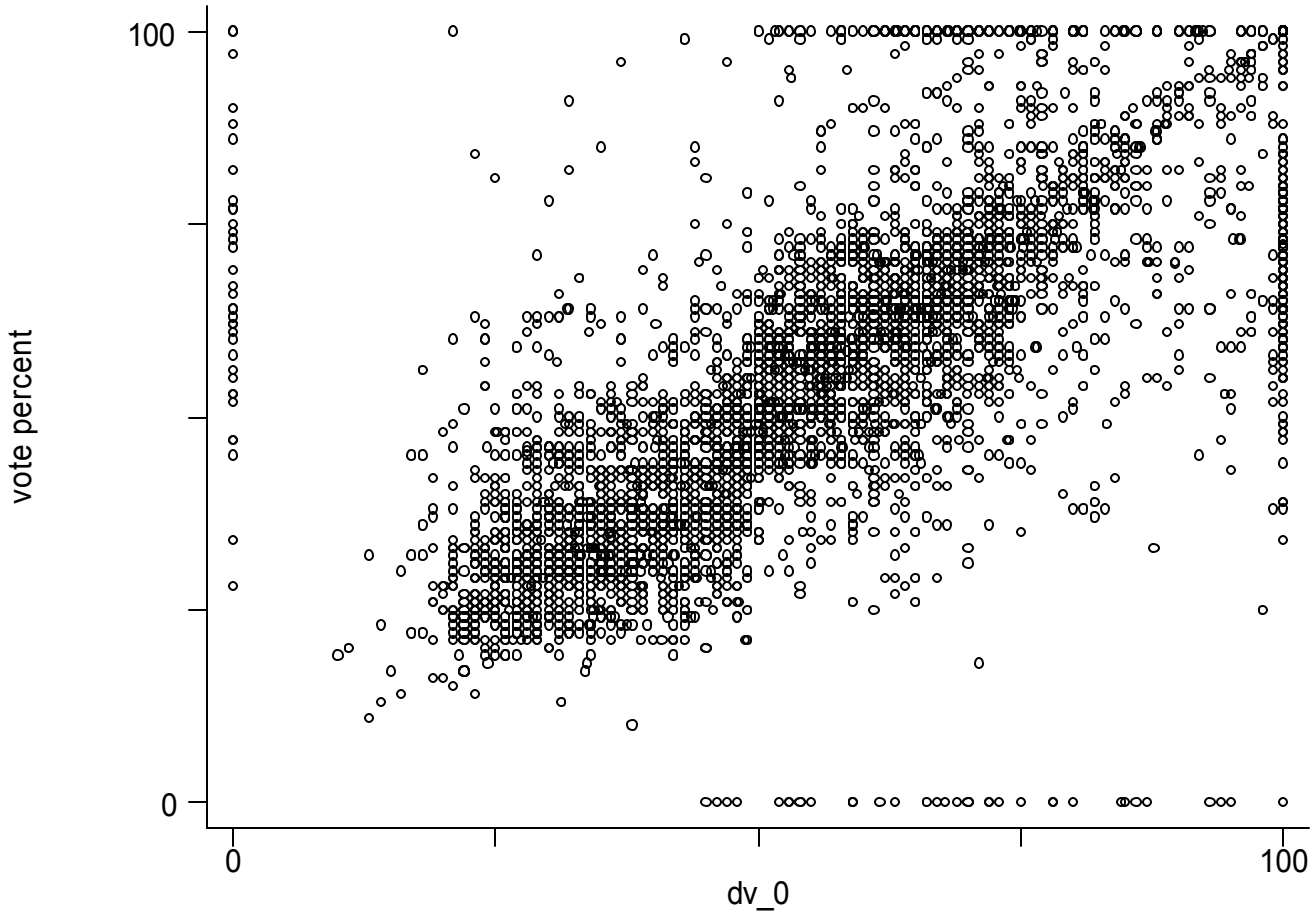
Suppose there are two variables, Y and X, each of which takes values 0 and 1. The interior values in the table, $P_{11}, P_{10}, P_{01}, P_{00}$ show the joint density of X and Y. The univariate density shown on the margins of the table: $P_{0.} = P_{00} + P_{01}$, $P_{1.} = P_{10} + P_{11}$, $P_{.0} = P_{00} + P_{10}$, and $P_{.1} = P_{11} + P_{01}$. These are called the *marginal* probabilities.

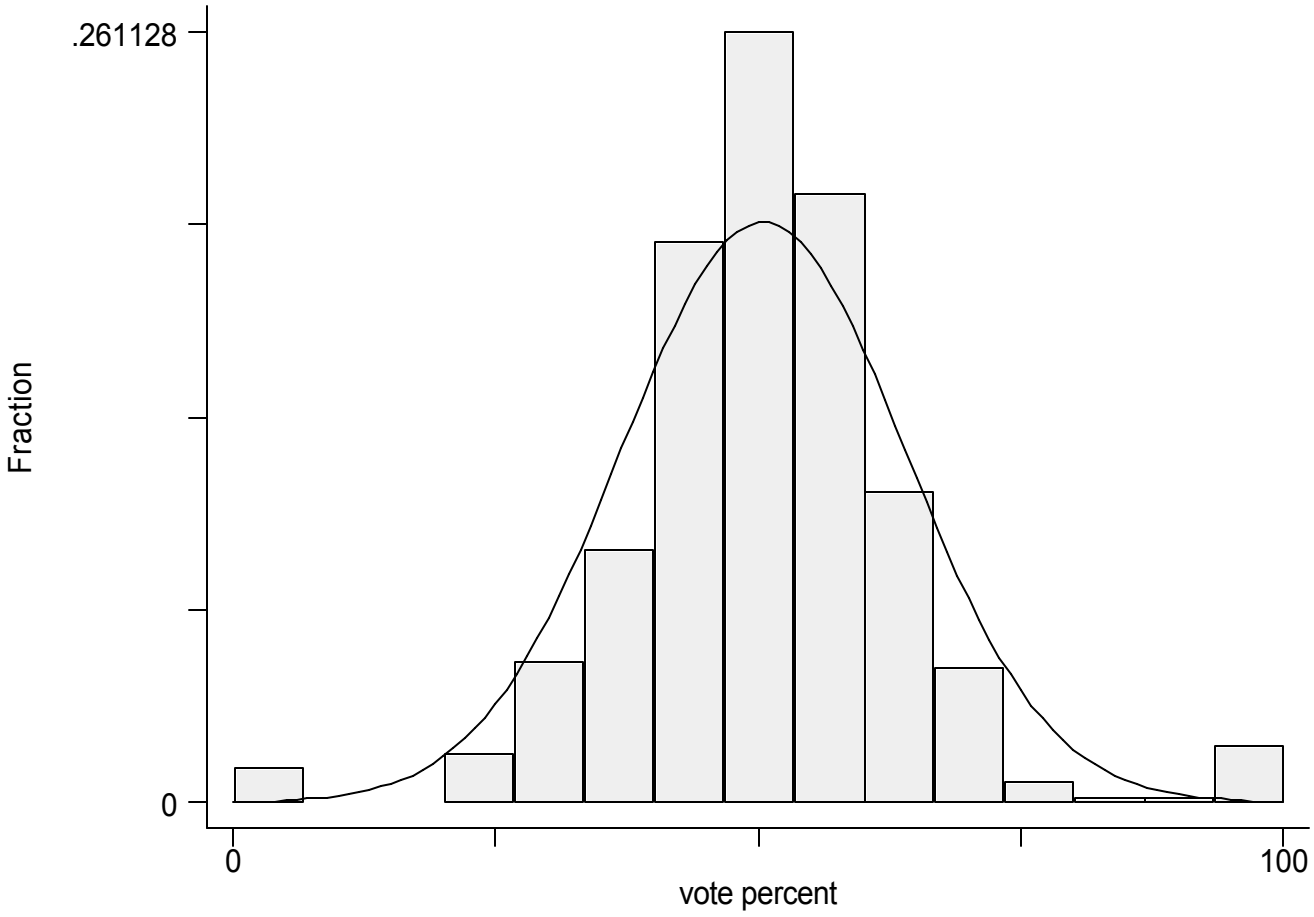
	X		
Y	0	1	
0	P_{00}	P_{01}	$P_{0.}$
1	P_{10}	P_{11}	$P_{1.}$
	$P_{.0}$	$P_{.1}$	1

From the joint distribution we can calculate the conditional probabilities, i.e., $P(X = x|Y = y)$ and $P(Y = y|X = x)$, using the multiplication rule. Conditioning means restricting the sample space. So, $P(X = 1|Y = 1) = P_{11}/(P_{10} + P_{11})$.

Case 2. Continuous Random Variables.

Let's consider the example of the distribution of votes across two elections. See the attached graphs. The marginal distributions are the distributions of votes in each election. The joint distribution captures the association between vote across elections.


100
0
vote percent
0
100
dv_0



The *joint normal* distribution provides a framework within which we think about the relationship among continuous random variables. As with the univariate normal, the density function depends on parameters of location and spread. To describe association between the variables, only one new parameter is needed, the *correlation*.

Two variables, say X_1 and X_2, follow a joint normal distribution. Their density function looks like a bell-shaped curve in three dimensions. The exact formula for this distribution is:

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2+2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]}$$

The component in the exponential is the formula for an ellipse. Specifically, if we set the formula equal to a number and map the values of X_1 and X_2 we will get an ellipse. The center or location of the ellipse is μ_1, μ_2. The parameter ρ is the major axis of the ellipse. In statistical terms, $\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$ is called the *correlation*, where σ_{12} is called the covariance. The constant in front of the exponential term guarantees the the entire joint density sums to 1.

The marginal distribution of one of the variables, X_i, is arrived at by integrating over the values of the other variable X_j. The result is a proper density function, all of whose density sums to 1. Consider, for instance, X_1.

$$f(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2}$$

The parameters σ_1 and σ_2, then, describe the (univariate) spread of each of the variables.

When we think about prediction and effects, we will typically study conditional relationships. To derive the conditional distribution, say, $f(x_2|x_1)$, we divide the joint probability density by the marginal probability density of X_1. Here we are just following the definition of conditional probability. That yields (after some algebra)s

$$f(x_2|x_1) = \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}} e^{-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(x_2-(\mu_2-\frac{\sigma_{12}}{\sigma_1^2}\mu_1)-\frac{\sigma_{12}}{\sigma_1^2}x_1\right)}$$

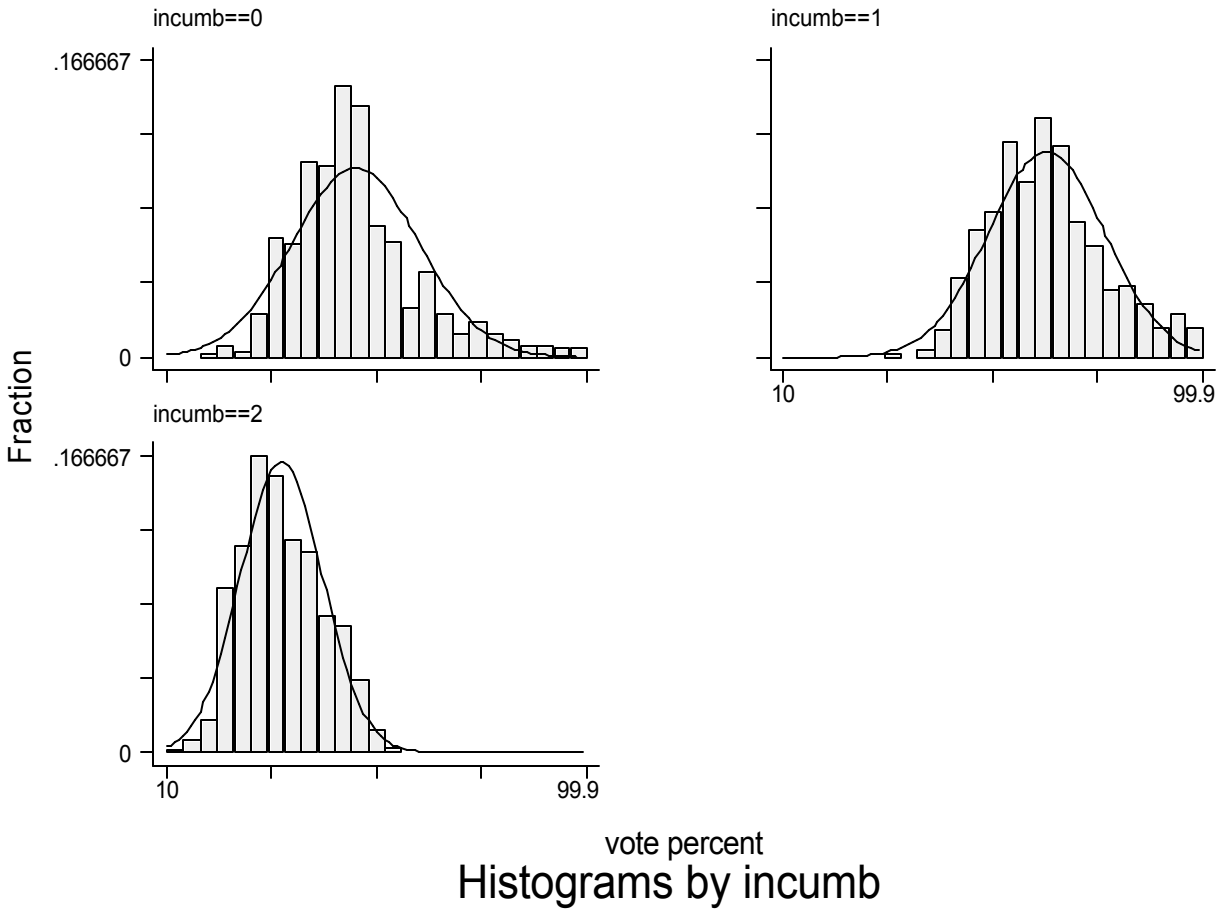
This has the form of a normal distribution where $\mu_{2|1} = \frac{\sigma_{12}}{\sigma_1^2}\mu_1) + \frac{\sigma_{12}}{\sigma_1^2}x_1$ and $\sigma_{2|1} = \sqrt{\sigma_2^2(1-\rho^2)}$.

The conditional distributions are a set of distributions, depending on the value of the conditioning variable X_j. In the case of jointly normal data, the spread parameters of the conditional distributions are always the same, $\sigma_{2|1} = \sqrt{\sigma_2^2(1-\rho^2)}$. The central locations of the conditional distributions are a linear function of the values of the conditioning variable. That is, $\mu_{2|1} = \alpha + \beta x_1$, where $\alpha = \frac{\sigma_{12}}{\sigma_1^2}\mu_1)$ and $\beta = \frac{\sigma_{12}}{\sigma_1^2}$.

Case 3. A Mix.

We often want to study problems that involve continuous and discrete random variables. For example, in US elections an important distinction is between elections where an incumbent is running and elections where a seat is open. This is a discrete random

variable. Researchers want to know how the vote depends on whether an incumbent runs. The random variables and data are readily displayed as separate distributions or graphs (as shown). In fact, the data are from a single joint distribution of an indicator variable (incumbency status) and a continuous variable vote. The graphs distinguish when a seat is open (incumb=0), when a Democratic incumbent runs for reelection (incumb=1), and when a Republican incumbent runs for reelection (incumb=2).



Density Estimates The histogram is an estimate of the density function. The histogram is constructed by dividing the range of the random variable into k bins of equal width (usually between 10 and 50). The frequency of observations in each bin consists of the estimated density associated with each interval. In STATA you can make a histogram using the command **graph x**. You control the bin-width using the option, i.e.: **graph x, bin(25)**.

A more sophisticated approach is Kernel Density estimation. Kernel estimation fits a continuous function to the data. There are a variety of functions one may fit. For presentation purposes kernel density estimates are more appealing that histograms

F. Expected Values

Probability distributions provide a powerful method for describing randomness (and chaos). In research and life, we most often want to eliminate randomness and focus on what is systematic and predictable. In statistics, we reduce randomness by averaging. By now you are familiar with the average value of a set of numbers. Analysis of problems using random variables requires a more general formulation of averaging, expected value.

This quantity has two interpretations. First, if realizations of a random variable (say an experiment) were observed a very large number of times, what is the average outcome you would observe across those many replications. Second, what is the "certainty equivalent" value of a random variable if there were no randomness. At what point is a risk neutral person indifferent between a specific value and the outcome of the random variable?

Set aside statistical questions, for a moment, and consider a simple example.

A lottery pays \$1,000 if you win and \$0 if you lose. You must pay \$1 to play, regardless of the outcome. Suppose every individual has an equal chance of winning a lottery and 10,000 play the game. The probability that you win is 1/10000. How much money do you expect to win? We calculate this using the idea of the certainty equivalent discussed in the first part of the course. (1/10000)\$1,000 + (9999/10000)\$0 = \$.10. However, you have to pay \$1 to play, so any individual expects to lose \$.90. Of, course no one actually loses \$.90; rather that is the certainty equivalent. That many people gamble suggests that they prefer the risk.

From the perspective of the state or the casino, they expect to earn \$10,000 from tickets sold, but must pay \$1,000 for the prize. We can arrive at the state's returns another way. They expect to earn \$.90 from each gambler; with 10,000 gamblers, that totals \$9,000 = \$.90×10,000.

F.1. Definition

Expected value is the average weighted value of a random variable where probabilities

serve as weights. It is defined as:

$$E(X) = 1P(X = 1) + 2P(X = 2) + 3P(X = 3)... = \sum_i iP(X = i)$$

$$E(X) = \int xf(x)dx.$$

This is usually called the mean value of X and denoted μ. It is the population or theoretical average value.

We may extend this definition to any function of the values of X. Let $h(X)$ be a function of X. Then

$$E(h(X)) = h(1)P(X = 1) + h(2)P(X = 2) + h(3)P(X = 3)... = \sum_i h(i)P(X = i)$$

$$E(h(X)) = \int h(x)f(x)dx.$$

Example. Suppose X follows the uniform distribution on [0,1]. What is $E(X)$? We can calculate the average value from simple geometry to be .5. We can also integrate:

$$\int_0^1 xdx = \frac{x^2}{2}|_0^1 = \frac{1}{2} - 0 = \frac{1}{2}$$

What is $E(X^2)$?

$$\int_0^1 x^d x = \frac{x^3}{3}|_0^1 = \frac{1}{3}$$

Example. Suppose X follows the Binomial distribution. One can prove, after some algebraic manipulation, that $E(X) = \sum_{k=0}^n \binom{n}{k} p^k(1-p)^{n-k} = np$. (I leave this as an exercise. The algebra takes some care.)

Statistics focuses on a particular set of expected values, called moments. The *moments* of a distribution consist of the expected values of all powers of a random variable. Let X be a random variable and k be integers 1, 2, 3, 4 .... The moments of the random variable are $E(X^k)$. That is, $E(X)$, $E(X^2)$, $E(X^3)$, $E(X^4)$, ...

We can readily estimate the moments using averages. If we observe a sample of n observations, then our estimate of $E(X^k) = \frac{1}{n}\sum_{i=1}^{n} x_i^k$.

A second sort of moment is "mean-deviated" – e.g., the second moment around the mean, the third moment around the mean. This is written $E[(X-\mu)^k]$. The second moment around the mean is the variance; the third moment around the mean is the skew; the fourth moment is the kurtosis. Higher moments do not have names.

The moments are of interest for two reasons. First, they summarize the data. The moments of a distribution contain the same information about the random variable as the density or probability function. One way to understand why this is the case is Taylor's Theorem. Using Taylor's Theoreme, any function can be written as a polynomial of the independent variable, where the constants on each polynomial equal the function evaluated at an arbitrary point. More properly we can summarize a distribution using the moment generating function, as discussed in Bulmer. This is an advanced topic in probability theory and one beyond the scope of our course. However, the basic point is important because it justifies using averages data and averages of squares, cubes, etc. of data to summarize probability.

More simply, you can think about the information contained in the mean, the variance, the skew, and the kurtosis. For the normal distribution the first moment equals the paramter μ and the second moment is the parameter σ^2. The Third Mean Deviated moment equals 0. The Fourth Mean Deviated Moment equals 3. Note: these are theoretical quantities. A test of normality is skew equals 0 and kurtosis equal 3. Consider the distribution of presidential vote across congressional districts. The mean is normalized to equal 0. The variance is .011 (about 1 percentage point). The skew is 1.06 and the kurtosis if 4.96. The positive number for the skew means that the tail stretches out to the right. The kurtosis larger than 3 means that there are more observations out in the tail than predicted by the normal distribution.

Second, moments offer an approach to estimation. The Method of Moments proceeds in three steps. (1) Express the moments of a density in terms of the parameters of the function. For example, for the normal distribution, $E(X) = \mu$ and $E(X^2) = \sigma^2 + \mu^2$. (We

could continue with more moments, but for this problem 2 is enough.) (2) Substitute the estimated moments for the theoretical moments. Continuing with the example, $\frac{1}{n}\sum x_i = \hat{\mu}$ and $\frac{1}{n}\sum x_i^2 = \hat{\sigma^2} + \hat{\mu^2}$. (3) Solve the equations expressing the estimated parameters solely in terms of the data. In the case of the normal distribution, we get $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_i x_i$ and $\hat{\sigma^2} = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$.

F.2. Mean and Variance

Two expected values are central to statistics – the mean and the variance. The mean is the expected value of X and is commonly denoted using the Greek letter μ, i.e., $\mu = E(X)$. The variance is expected (squared) distance of any value of X from the mean and is denoted with the Greek letter σ^2, i.e., $\sigma^2 = E[(X - \mu)^2]$.

These two important parameters capture the central tendency and the spread of the random variable or data. In other words, the mean describes what is predictable and the variance describes what is chaotic–the yin and yang of statistics.

There are other measures of central tendency, most notably the median and and mode. The median is the 50th percentile value; the mode is the value of X with the highest frequency. For a symmetric distribution, such as the normal, these will be the same as the mean. There are also other measures of shape and spread that use higher ordered mean-deviated moments: $E[(X - E(X))^3]$, $E[(X - E(X))^4]$,etc. The third-mean deviated moment is called the skew. It measures the extent to which the data have a disproportionately high frequency of values above (or below) the mean. Income distributions, for example, are highly skewed. The fourth mean-deviated moment is called the kurtosis. It measures whether there are a high frequency of values in the "tails of the distribution." The mean and the variance, as explained below, are sufficient to characterize most problems of inference.

1. Mean Values

The theoretical mean is the weighted average of the values, where weights are probabil-

ities. If the distribution is skewed the mean will lie above the median, which will lie above the mode. We must distinguish the sample mean $\bar{x}$, which is the simple mean of the n observations in a study, from the population mean. The population or theoretical mean is a number; it is a fixed quantity that we may or may not know. The sample mean is the sum of n random variables and this is itself a random variable with a probability distribution.

Random variables are often transformed, which alters their mean values. The most common transformations are linear, and there are two sorts.

- Expected Values of Linear Functions are Linear Functions of Expected Values:

$$E(a + bX) = a + bE(X).$$

As a demonstration, consider a continuous random variable X.

$$E(a + bX) = \int_{-\infty}^{\infty} (a + bxf(x))dx = \int_{-\infty}^{\infty} af(x)dx + \int_{-\infty}^{\infty} bxf(x))dx$$
$$= a\int_{-\infty}^{\infty} f(x)dx + b\int_{-\infty}^{\infty} xf(x))dx = a + bE(X)$$

- Expectations of Sums of Random Variables are Sums of Expectations. If X and Y are random variables, then

$$E(Y + X) = E(X) + E(Y)$$

Assume the variables are continuous with joint density $f(x, y)$.

$$E(Y + X) = \int\int (y + x)f(x, y)dxdy = \int xf(x, y)dxdy + \int yf(x, y)dxdy$$
$$= \int xf(x)dx + \int yf(y)dy = \mu_x + \mu_y$$

These results are immediately helpful in analyzing the mean of random variables that are themselves sums. Take for example, the Binomial distribution. The Binomial is the sum of n independent Bernoulli trials. Using the addition rule above allows a simple computation of the Binomial mean. $E(X) = E(X_1 + X_2 + ... + X_n) = p + p + ... + p = np$. Likewise, the mean of the Poisson is λ, since $\lambda = np$. When calculating the normal approximation to the Binomial we use $\mu = np$.

2. Variances

The variance is the average squared distance of observations from the data. The Standard Deviation is the Square Root of the Variance. The standard deviation and the mean are used in "normalizing" data. The standard deviation measures the "units" of the variable.

Example. X is distributed Uniform(0,1).

$$V(X) = \int_0^1 (x - .5)^2 dx = \int_0^1 (x^2 - x + \frac{1}{4}) dx = \frac{x^3}{3} - x + fracx4|_0^1 = \frac{1}{12}$$

Variances of linear functions and sums are essential in statistics.

- $E[(X - \mu)^2] = E(X^2) - \mu^2$.

  $E[(X-\mu)^2] = E[X^2-2X\mu+\mu^2] = E[X^2]-2E[X\mu]+\mu^2 = E[X^2]-2\mu^2+\mu^2 = E[X^2]-\mu^2$

- Jentzen's Inequality.

  $$E(X^2) > E(X)^2$$

  More generally, for any convex function $g(X)$, $E(g(X)) > g(E(X))$.Note: This fact guarantees that variances are always positive. This fact is also very useful in thinking about risk analysis, as $g(\cdot)$ can be treated as a utility or loss function.

- $V(a + bX) = b^2 V(X)$.

  $V(a+bX) = E[(a+bX-a-b\mu)^2] = E[(bX-b\mu)^2] = E[b^2(X-\mu)^2] = b^2 E(X-\mu)^2 = b^2\sigma^2$

- If X and Y are independent $V(Y + X) = V(X) + V(Y)$

These rules make easier to calculate variances of random variables that are sums of random variables. Consider the Binomial example. $V(X) = V(X_1 + X_2 + ... + X_n) = V(X_1) + ... + V(X_n) = p(1-p) + ... + p(1-p) = np(1-p)$

Of particular interest is the sample mean. What is the variance of the sample mean? The sample mean is a random variable, because it is a sum of random variables. We determined

that the expected value of the sample mean is μ. If the observations are independent of one another (or nearly so, as with samples that are small relative to the population), then we can calculate the variance of the sample mean using the addition rule for variances.

$$V(\bar{x}) = V(\frac{1}{n}\sum_i x_i) = \left(\frac{1}{n}\right)^2 V(\sum_i x_i) = \left(\frac{1}{n}\right)^2 \sum_i \sigma^2 = \frac{\sigma^2}{n}$$

We have assumed two things in this derivation, both enter in the second to last equality. First, the observations are independent; hence, $V(x_1+x_2+...x_n) = V(x_1)+V(X_2)+...V(x_n)$. Second, the variances are the same for all of the cases. That is the observations are drawn from the same population. Violations of either assumption, as occurs, for example, with sampling without replacement, would lead us to alter this formula somewhat.

The interpretation of the two results we have derived about the sample mean are quite interesting. Were we to conduct the same study repeatedly and estimate the population mean using the sample mean, the sample mean would vary in a predictable way. The central tendency of the distribution of sample means would be the population mean, μ, and the rate of variation from sample to sample is $\frac{\sigma^2}{n}$. These two features of the data are the foundation for most of statistical inference and design. We can uses these facts to figure out the optimal sample size for a study and the likelihood of making a disparate observation. We also learn from these facts that we don't have to do studies repeatedly in order to draw inferences, a single large survey may do. We will build these ideas in the next section of the course.

Before we turn to statistics proper, I will conclude the discussion of mean and variance with one important result and one important idea about means and variances. The important result is Chebychev's inequality.

F.3. Confidence Intervals, Chbebychev's Iequality, and the Law of Large Numbers.

Consider the accuracy of the sample mean as an estimate. How close is it to the right answer? Suppose that we have calculated the sample mean as an estimate of the population mean. The sample mean does not equal the population mean, unless we are very lucky. Rather sample means vary randomly around the true mean. The calculation we have just

made tells us the rate of variation is $\frac{\sigma^2}{n}$. What is the probability of observing a sample mean that is far from the true mean? That is, if k is a number that we deem "far from μ, what is $P(|\bar{x} - \mu| > k)$?

We may ask this question more generally for any random variable X. What is the probability of observing a large deviation from the mean? The mean and the variance are powerful tools because they can be used to bound the variation in a random variable. Chebychev's inequality states that the variance tells us the maximum amount of density within an interval around the mean *for any distribution.*

Consider a deviation from the mean $|X - \mu|$. What is the probability of observing a deviation from μ at least as large as a specific number k? Chebychev's inequality provides a general formula for this bound, regardless of the underlying distribution:

$$P(|X - \mu| > k) < \frac{\sigma^2}{k^2}$$

That is, the probability of observing a deviation from the mean at least as large as k is less than the variance of X divided by k^2. Chebychev's inequality can be rewritten if we let $t\sigma = k$, where t is a number. Then,

$$P(|X - \mu| > t\sigma) < \frac{1}{t^2}$$

Reformalated this way, Chebychev's inequality is interpreted as follows. The probability of observing a value of X more than t standard deviations away from the mean is less that $1/t^2$. This bound may not be terribly tight for some distributions, but it is generally useful.

Let us consider two sorts of problems.

First, how wide an interval around the mean guarantees that the probability of a large deviation is less than 5 percent? This is a problem of constructing a *confidence interval.* We begin with a probability statement about our desired level of confidence. We wish to be highly confident (i.e., at least 95 percent confident) that the population mean lies in an interval. Given the desired level of confidence how wide must the interval be?

Chebychev's inequality tells us what the widest possible interval is. Set the probability of a deviation of at least $t\sigma$ to be less than .05. That is, the probability weight in the "tails"

of the distribution must be smaller than .05 The second version of Chebychev's inequality can be used for this problem. Set the value of the bound $\frac{1}{t^2} = .05$. Solving this equation, $t = \sqrt{20} \approx 4.5$. In other words, 4.5 standard deviations away from the mean guarantees at least .95 probability of covering the population mean. If we know or estimate σ and we observe a value of x, then we can calculate the widest 95 percent confidence interval as $x + / - 4.5\sigma$.

Continuing with this problem, we can ascertain the widest possible confidence interval for the sample mean. What is the interval around the sample mean that one could construct so as to be sure of covering the true mean with probability .95. We said that the variance of $\bar{x}$ is σ^2/n. The standard deviation of $\bar{x}$ is $\frac{\sigma}{\sqrt{n}}$. Hence,

$$P(|\bar{x} - \mu| > 4.5\frac{\sigma}{\sqrt{n}} < .05$$

In other words, we are at least 95 percent sure that the interval $\bar{x} + / - 4.5\frac{\sigma}{n}$ includes the population mean μ.

When public opinion pollsters report data on presidential approval, vote intentions, and other features of public attitudes, then often report a margin of error. The margin of error is the half-width of the 95 percent confidence interval, say +/- 3 percentage points.

For most statistical problems, we will learn below, we can improve on Chebychev's inequality, because we know more about the distribution.

Second, how close is the sample mean to the right answer? The first formulation of Chebychev's inequality reveals that we can be arbitrarily close to μ. Again substitute the formulas for the sample mean and variance of the sample mean into Chebychev's Inequality:

$$P(|\bar{x} - \mu| > k) < \frac{\sigma^2}{nk^2}$$

The term on the right-hand side depends on a parameter σ^2, a constant k, and the sample size n. We can increase the sample size to make the term on the right hand side arbitrarily small. As n becomes infinitely large the term on the right hand side approaches 0. Also, we can choose k to be any arbitrarily small positive number. Hence, by increasing the sample

size we can squeeze to 0 the probability that $\bar{x}$ deviates from μ by more than an arbitrarily small number.

This limiting result is a special case of the Law of Large Numbers. Any statistic that we calculate that takes the form of an average will be arbitrarily close to the parameter that the statistic estimates when we have large samples. Generally, this concept is called the *consistency* of a statistic.

An important version of the Law of Large Numbers is the Binomial. Suppose that we wish to study the population frequency of a certain trait or the probability of an event, p. We collect a sample of size n with replacement from the population. The ratio k/n estimates p, and k follows the Binomial distribution. The variance of k/n is $\frac{p(1-p)}{n}$. [Question: Why?] What is $P(|\frac{k}{n} - p| > \delta)$? We know from Chebychev's Inequality that this probability is smaller than $\frac{p(1-p)}{n\delta}$. As n grows the bound approaches 0, and the probability that k/n deviates from p by more than a tiny amount is 0.

The Law of Large Numbers brings us back to our definition of probability. We said that there are two common definitions of probability – subjective belief and long-run frequency. We now have a clear and general definition of long-run frequency. Consider what the Binomial result means for a coin tossing problem. Each time we toss a coin the probability of heads is p and of tails $1-p$. The true probability is never observed because the sample space is infinite; one could toss the coin forever and never enumerate all possible coin tosses. However, we might feel that after 100 or 1000 coin tosses the observed frequency is close enough to the true probability. Chebychev's inequality, then, implies that frequencies approach the true probability of an event, even when the population or sample space is infinite.

F.4. Mean Squared Error

In statistics there is another sort of Variance, called the mean squared error. Mean squared error is the average or expected error in a statistic. Mean squared error encompasses two very important measurement concepts, bias and precision.

Statistics are quantities that we can compute using data and that summarize the data,

just as the moments summarize the distribution. The average value, $\bar{x}$, is the most frequently used statistic as it measures central tendency. The estimated variance is $\hat{\sigma^2} = \frac{1}{n}\sum_i (x_i - \bar{x})^2$; the sample covariance is $\hat{\sigma_{xy}} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$;the sample skew is $\hat{\psi} = \frac{1}{n}\sum_i (x_i - \bar{x})^4$; the sample kurtosis is $\hat{\kappa} = \frac{1}{n}\sum_i (x_i - \bar{x})^4$. There are many other statistics, such as the median, the range, and the mean absolute deviation, that we might study as well.

Each statistic is taken as a measure of a parameter or theoretical relationship. The sample average estimates the population mean. A regression estimates a theoretical relationship of a dependent variable to one or more independent variables. Two important features of all statistics – their *bias* and *precision*. In general, we desire statistics that are unbiased and precise. An unbiased statistic is one that we expect will give the right answer; a precise statistic is one that will not vary much from measurement to measurement. Mean squared error summarizes both of these ideas.

Let us call any statistic $T(X)$, a function of the data. Let us suppose that $T(X)$ is constructed so as to estimate θ. $T(X)$ is a random variable because it is a function of random variables. It will vary from sample to sample. For example, suppose we conduct a survey of 1000 people to ascertain their attitutdes toward the sitting president. We will get one estimate of the level of support for the president. If we were to conduct a second survey using the identical sampling procedures we will get another sample and another, different estimate because the people in the first survey differ from those in the second survey. We could imagine repeatedly sampling and map out a frequency of estimated levels of support. This is called the *Sampling Distribution*.

The bias of a statistic is defined as the expected deviation of the statistic from the parameter that the statistic estimates. Assume that $E[T(X)] = \tau$. We write Bias$= E[T(X) - \theta] = \tau - \theta$. The precision of the statistic is the variance of $T(X)$. ; so $V(T(X)) = E[(T(X) - \tau)^2]$.

Mean squared error is defined as $E[(T(X) - \theta)]$. We can write this in terms of bias and precision:

$$E[(T(X) - \theta)^2] = E[(T(X) - \tau + \tau - \theta)^2] = E[(T(X) - \tau)^2 + 2(T(X) - \tau)(\tau - \theta) + (\tau - \theta)^2]$$

$$= V(T(X)) + (\tau - \theta)^2$$

A general goal of statistics is to improve our design and estimation so as to minimize mean squared error. We will typically be interested in unbiased statistics and then try to minimizee the variance (or noise) of the statistic. Usually, we will discard unbiased or inconsistent statistics, so much quantitative work begins with attempts to eliminate bias from studies.

When repeated measures are available we can measure mean squared error directly as the variance across measures. For example, if we have 10 surveys conducted using the same general methodology and questions one measure the variance across the surveys to see if there are biases due to, say, different firms. Also, this technique is used to callibrate interviewers or testers. The measure used is just the estimated variance across surveys. Let $\bar{\bar{x}}$ be the average of the sample averages and $\bar{x}_j$ be the sample average of one survey. Then the estimated MSE is

$$\frac{1}{J-1}\sum_{j=1}^{J}(\bar{x}_j - \bar{\bar{x}})^2$$

F.5. Covariance, Correlation, and Reliability of Measures.

When we deal with joint data, the covariance will summarize the joint relationship. Covariance is defined:

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

Sometimes we denote $Cov(X,Y) = \sigma_{xy}$. The covariance depends on the units, which makes for difficult interpretation when two variables are in different units. The standardized covaraince is called the correlation and denoted using the Greek letter ρ.

$$\rho = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = Cov\left(\frac{X}{\sigma_x}, \frac{Y}{\sigma_y}\right)$$

Several important features of covariance are useful in algebraic manipulation:

- Covariance is defined as $Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = E(XY) - \mu_x\mu_y$.
- $Cov(a+bX,Y) = bCov(X,Y)$.
- If X and Y are independent, then $E(XY) = E(X)E(Y)$. Hence, if X and Y are independent $Cov(X,Y) = 0$.
- Generally, $V(X+Y) = V(X) + V(Y) + 2Cov(X,Y)$.

$$V(X+Y) = E[((X+Y)-(\mu_x+\mu_y))^2] = E[(X-\mu_x)^2+(Y-\mu_y)^2+2(X-\mu_x)(Y-\mu_y)]$$

$$= V(X) + V(Y) + 2Cov(X,Y).$$

- Cauchy-Schwartz Inequality. $E(X^2)E(Y^2) > E(X)^2E(Y)^2$. This guarantees that the correlation always lies between 1 and -1.

We will use covariance to study relationships among variables later in the course. Now I would like to consider them as providing a third important measurement concept in addition to bias and precision. A statistic or measurement is said to be *reliable* if repeated measurements are highly correlated. This idea is very widely applied in educational testing and in the construction of indexes.

Within political science an important debate has concerned the stability of public opinion. Converse (1964) argues that public opinion is not very stable or well-grounded. He looked at three waves of a panel study and correlated answers to the same questions by the same people in three successive elections (1956, 1958, and 1960). The results were not happy. The correlations were quite low, typically in the range .3 to .5. From this Converse concluded that the public does not hold coherent ideologies or beliefs.

Chris Achen (1975) critiques Converse's important conclusions as reflectin the low reliability of the survey questions, more than instability in people's beliefs. Achen argues that responses to survey items reflect systematic components, people's true beliefs, and noise due

to poor question wording, also known as measurement error. People try to guess what the question means, leading to random variation in responses. Let u be the survey response error and X^* be the true belief. Then, $X = X^* + u$. We can measure responses at different times. Assuming the measurement error is unrelated to the true attitude, the covariance between the survey responses at two different times will be $Cov(X_1, X_2) = Cov(X_1^*, X_2^*) + Cov(u_1, u_2)$. Assume the measurement error is unrelated. The Variance of the question responses at each time will equal $V(X_1) = V(X_1^*) + V(u_1)$ and $V(X_2) = V(X_2^*) + V(u_2)$. The true correlation in attitudes is $\rho_{12}^* = \frac{Cov(X_1^*, X_2^*)}{\sqrt{V(X_2^*)V(X_1^*)}}$ The correlation between survey questions across the two times is:

$$\rho_{12} = \frac{Cov(X_1, X_2)}{\sqrt{V(X_1)V(X_2)}} = \frac{Cov(X_1^*, X_2^*)}{\sqrt{(V(X_2^*) + V(u_2))(V(X_1^*) + V(u_1))}} < \rho_{12}^*$$

Achen further argues that the true correlation in attitudes is plausibly as high as .8 to .9 on most items. The problem he concludes is low reliability in the instrument used to measure attitudes.

Another important debate within political science concerns the measurement of the preferences of members of Congress. Various roll call voting scores have been constructed by interest groups, such as Americans for Democratic Action, and by academics, such as Poole and Rosenthal or Heckman and Snyder. ADA identifies 20 key roll call votes and constructs a "liberalism" measure based on the number of times that members of Congress vote with the ADA. The academic models consider the correlation between all legislators roll call votes on all issues and then try to uncover the number of dimensions and score legislatures on these dimensions. At least on the first dimension, the different methods produce indexes that are very highly correlated. The correlation between ADA scores and Poole-Rosenthal scores is typically around .90 to .95.

Part 3 Summary

So far in this course we have focused on issues that may be broadly thought of as mea-

surement. We have defined random variables and developed the key mathematical concepts with which to study random variables.

- A random variable is a variable whose values occur with a specific frequency. Properly speaking a random variable is a function which maps values into frequencies.
- A probability function defines the frequency with which each value of a random variable occurs. This function must satisfy three properties: (1) the probability of any event lies between 0 and 1, (2) the probability of at least one value occurring is 1, and (3) the probability of two disjoint events is the sum of the probability of each event.
- Expected value is the weighted average of all values of a function in which we treat the probabilities as weights.

As examples we have considered many statistical problems, such as maximimum likelihood and method of moments estimation and confidence intervals. In the next part of the course we will use these elements as building blocks of statistical modeling.