INTRODUCTION TO STATISTICS FOR POLITICAL SCIENCE:
Part 4: Statistical Modelling

Stephen Ansolabehere
Department of Political Science
Massachusetts Institute of Technology

Fall, 2003

Part 4.
STATISTICAL MODELS AND METHODS

Any study of a problem begins with a clear statement of what we seek to learn and why we seek to learn it. Most academic studies have in mind several competing explanations of a phenomenon of interest, and studies begin by deriving from the competing explanations specific expectations about the behavior in question. These questions predict patterns of behavior or relationships between variables. Empirical researchers attempt to build statistical models that are appropriate to the theoretical questions at stake and to draw inferences about theories by testing predicted outcomes against observed data.

Let us consider a couple of established areas of research in political science.

Why has turnout declined over the last 40 years in US elections? One common approach is to study who is more likely to vote and to measure the effects of demographic characteristics and political attitudes on participation. Several very important factors appear, such as age, education, and income. Older, better educated, and wealthier people participate more. Curiously, these have all increased since 1960 (the high-water mark of turnout in modern US elections), but turnout has declined. What might explain changes in turnout?

What is the incumbency advantage and what factors contribute to it? Over the last 50 years, the reelection rates and vote margins of US House incumbents have grown dramatically. This is a distinctive feature of American elections, and it is a challenge to know why it occured. Among the conjectures are that the technology of communication changed with the introduction of television and that the rules of politics shifted with the imposition of new redistricting rules in the 1960s. Also, it is claimed that interest groups caused the growth of the incumbency advantage because interest groups give money to politics, that the primary elections caused the incumbency advantage, and that the decline of political party organizations contributed to the rise of personal politics.

What are some of the key variables at stake and how might we try to assess their importance? How do we measure the incumbency advantage? How can we study the causes of the incumbency advantage? What data would help to address the causes of the incumbency advantage?

These two topics in political science have been subject to intense statistical scrutiny. They each began with observations of fact: declining ratio of votes cast to the number of people in the voting aged population, and increasing reelection rates and vote margins of House incumbents. In each, there has been robust debate over measurement questions; there have been advances in the measurement and modeling of the variables in question; and, most importantly, there has been a cumulation in knowledge.

Consider for example the study of turnout. Measuring the turnout rate has long been problematic, as the baseline is difficult to establish. Popkin and Rabinowitz have recently argued that there has been relatively little decline in voting in the US because of the growth of immigration. Establishing the causes of turnout has been somewhat easier. Verba and Nie established the main sociological predictors of participation, especially age and education. Brody, however, notes that these cannot cause declining participation. And Rosenstone and Hansen find using data from the National Election Study from 1948 to 1992 that declining turnout is attributable to declining party electoral activity. Also, comparative political studies of turnout (such as Powell) show that electoral systems with PR have much higher turnout. After 30 years of intensive study of the subject we know what are the strongest predictors of participation and we have a new conjecture that political organizations may be responsible for the decline in turnout. We don't yet understand the psychology of voters as it relates to participation. Why do better educated people vote more?

The process of model building in these areas consists of the search for a simple set of explanations for behavior. What is an adequate explanation? Presumably one that explains a high fraction of the variation in behavior and that predicts behavior very well. In reaching these standards it is clear that we need rules of scientific evidence. How do we know when we have made an improvement over past studies? What are appropriate and inappropriate ways to analyze data? Can others replicate an analysis? Can they replicate a finding using identical methods? What would be the ideal study, and how would we implement it?

These are all problems of estimation and inference, and ultimately design. In this section of the course we will develop statistical models and methods by building up from simple problems to more complicated ones.

1. General Concepts of Statistical Modeling

In this section, we develop the general concepts for statistical modeling using a simple Bernoulli example. The goal is to develop the ideas of data summary, estimation, and inference using a very simple problem. We will complicate this framework as we consider more complex questions and study designs.

To give the subject some empirical flesh, consider the following problem. The Los Angeles County Recorder and Registrar maintains the voter registration lists. The county has an estimated 5.5 million eligible voters and 4.0 million names on the voter registration lists. The voter registration lists may contain many duplicate registrations or obsolete registrations because people move within the county or leave the county. Michael Alvarez and I conducted a study designed to increase turnout. One part of this study involved an attempt to measure obsolete registration listings, so that we could gauge what fraction of the population was actually registered and what fraction of truly registered voters voted. We randomly selected 25 precincts out of 4,922. Within each of these precincts (of about 400 people each) we randomly selected 100 people. We then mailed two pieces of first class mail to each of the listings on the sample. On the envelope were explicit instructions to return the mailing if the person to whom the letter was sent no longer resided at the address. Because the mail was sent first class all undeliverable mail was returned. What fraction of registrations on the LA County Registrar's list are obsolete?

There is a population fraction of obsolete listings on the registry; we denote this fraction as p. Once we estimate p we can calculate the estimated fraction of registered voters who voted. In the 2002 election, 1,784,320 people voted – 44.8 percent of the names on the registry. To calculate the actual percent of registered people who voted, we need to adjust the baseline number of names on the registry. The actual percent who voted is: $44.8/(1-p)$.

What is p?

1.A. Data Summary

We begin by specifying how the data were generated. It is useful to distinguish between two sorts of studies you will conduct and encounter – "designer data" and "found data." Roughly the distinction is this. Many studies, such as surveys and lab experiments, are carefully designed. The researchers choose the content, such as the questions, and the sample sizes. These choices are made with specific theoretical conjectures in mind, subject to budget constraints. In many ways these are the ideal studies described in your statistics books. I think the majority of studies consist of "found data." Researchers either analyze data collected for some other reason or data that nature generated, such as the historical record of elections, stock returns, or wars. With found data you get what you get. Found data sounds like it has a bit of mutt and mongrel to it. It does, and this is most of what we do in social sciences. The question is can we figure out how to make the most of this information and to avoid pitfalls of improper inference from data? Of course, the same is true of designer data. When we have the opportunity to design a study, we want to design the best study possible. I will usually treat data as if we had designed the study. The same thinking goes into "found data."

When we write about research we must be very clear about the data we have at hand – what are its strengths and weaknesses, how does it improve on or supplement other empirical studies? Most studies present the "Data and Methods" toward the beginning of the presentation. It is good to present what is known about the data. For example, if a survey is used, what is known about the validity and reliability of the questions.

Statistics texts use a generic description of a study. A researcher makes n independent observations of a random variable $X_1, X_2, X_3, ..., X_n$. Sometimes this is stated as "a sample of n observations." Even for observational data, such as the incidence of wars over the last 100 years, data are a sample from the set of all possible occurences. We will typically assume independence as the default. It is possible that there is some dependence within the data, and this is interesting to model.

What is the probability function associated with these n random variables? The joint density of the data is called the *likelihood*. Let θ represent the parameters of the joint density. The likelihood is

$$L(\theta) = f(x_1, x_2, x_3, ...x_n; \theta)$$

I use the semicolon to separate the parameter of f from the values of the random variables.

Two important common assumptions about the generation of data are that each observation is *independent* and that the density functions for of the each observations are *identical*. We may, then, rewrite the likelihood as

$$L(\theta) = f(x_1; \theta)f(x_2; \theta)f(x_3; \theta)...f(x_n; \theta) = \Pi_{i=1}^{n} f(x_i; \theta)$$

The likelihood function is often transformed using logarithms, which makes the function linear and has an interpretation as an entropy function.

$$ln(L) = \sum_{i=1}^{n} ln(f(x_i; \theta))$$

Consider a Bernoulli random variable, $X = 1$ with probability p and $X = 0$ with probability $1 - p$. For example, I conducted a survey for Los Angeles County to measure the incidence of obsolete voter registrations. $X = 1$ means incorrect address. We randomly chose 25 precincts, out of 4,922. Within each of these we chose 100 persons to receive a first class mailing. The mailing was to be returned if the person was no longer at that address. The probability function for any one observation is,

$$f(X) = p^X(1 - p)^{1-X}.$$

Suppose we conduct a random sample survey without replacement from the population to measure X. Then we have n observations with the joint density:

$$L(p) = \binom{n}{x} p^x(1 - p)^{n-x}$$

and

$$ln(L) = ln\binom{n}{x} + xln(p) + (n-x)ln(1-p)$$

We may also summarize the data using the moments. Here we see a bit more clearly what identicality entails. The mean and variance of the distributions of each of the random variables are assumed to be the same (ie., from the same population): $E(x_i) = \mu$ and $V(X_i) = \sigma^2$. If each trial or draw from the population had its own mean and variance, these moments would depend on the individual case i.

Continuing with the Bernoulli: $E(X) = p$ and $V(X) = p(1-p)$.

1.B. Estimation

The two approaches to data summary give us two different approaches to estimation.

The method of moments involves using the sample statistics to estimate the mean. Once we estimate the mean, we can also, in this problem, estimate the variance. There is only one parameter.

The sample mean is

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{k}{n}$$

The *principle of maximum likelihood*, due to R. A. Fisher, involves choosing a value of θ that makes the probability of observing the data most likely. Specifically, $\hat{\theta}$ is a guess of the value of θ such that $L(\theta)$ is highest. This can be arrived at through the first derivative of the log-likelihood function.

$$\frac{\partial ln(L)}{\partial \theta}|_{\theta=\hat{\theta}} = 0$$

In the Bernoulli case,

$$\frac{\partial ln(L)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p}$$

Setting this equation equal to 0 and solving for $\hat{p}$ yields $\hat{p} = \frac{k}{n}$.

In the Los Angeles county data, 12 percent of the 2500 observations were returned. This is our estimate of the fraction of the registration roll that is no longer current.

1.C. Inference

Inference may be divided into two subjects. Statements of confidence in estimates and hypothesis tests. There is a close link between the two which we will develop here.

i. Confidence Intervals

The simplest sort of inference we make is to construct confidence bounds. What is the interval around $\hat{p}$ such that we are 95 percent confident that the interval covers the true proportion. That is, let us calculate

$$P(|\hat{p} - E(\hat{p})| > t\sqrt{V(\hat{p})}) \leq \alpha,$$

where α is a suitably small probability of a deviation from the mean, typically .05.

To make this calculate we need to understand three features of the distribution of $\hat{p}$ – the mean, the variance, and the appropriate values of t. Of course, t is determined by the distribution function of $\hat{p}$.

The mean and variance of $\hat{p}$ are straightforward. Assuming that the sample and the measurement are not subject to biases, $E(\hat{p}) = E((\frac{1}{n})\sum x_i) = np/n = p$. Also, assuming indepedence of observations, $V(\hat{p}) = V(\frac{1}{n})\sum x_i) = np(1-p)/n^2 = \frac{p(1-p)}{n}$. The square root of the variance of the estimate is called the *standard error*.

To make the probability calculation we could calculate the widest possible bounds using Chebychev's inequality. However, we can do a lot better. The statistic $\hat{p}$ will follow the normal distribution quite closely. Why? The estimate equals k/n. Since n is a number and k is random, we know that the distribution of $\hat{p}$ is determined by the distribution of $S_n = \sum x_i = k$.

The last result is a special case of the Central Limit Theorem. The Central Limit Theorem states that a sum of random variables will be distributed normally with mean $n\mu$ and variance $n\sigma^2$. Because most statistics are sums this means that almost all inference can be based on the Normal distribution, regardless of the underlying distribution of the data.

We may calculate the value of t from the Standard Normal probability table. To cover 95 percent of the distribution $t = 1.96$. That is, any observation more than 2 standard errors from the true p is likely to occur only 5 percent of the time or less. To cover 90 percent of the distribution requires $t = 1.645$.

Now, let us reconsider the LA County voter registration data. With a sample of 2500 people, what is the likely range of p? We calculate this as .125 plus or minus $\sqrt{(.125)(.875)/2500}$. This interval is approximately .11 to .14. We are 95 percent confident that the fraction of duplicate and obsolete registrations on the LA Country rolls is between 11 and 14 percent of all names.

ii. Hypothesis Tests

An hypothesis test begins with a conjecture or theory of behavior. The conjecture predicts that the data behave as if the underlying parameters of a function equalled some specific value. Common hypotheses are that two samples are identical or that there is no relationship among a set of variables. Suppose, for example, that a colleague has made a guess about the true error rate of .1 and used that in a research project. We can treate $p_0 = .1$ as an hypothesized value. Do the data support this assumption?

To construct the test, we must first consider what the possible outcomes of a test are. We will use data to reach conclusions. Hopefully, we reach the correct conclusions from the data – that the hypothesis is false when it is in fact false and that it is true when it is in fact true. However, we might make two sorts of errors with the data. We might judge the hypothesis to be false when it is not, or we might judge the hypothesis to be true when it is false.

Once we collect data we want to use the data to draw an inference about the hypothesis. Does the data cast doubt on the hypothesis or support it? Of course, we do not observe p_0 directly. In stead, we collect data and compare that data to what we think the data would look like were the hypothesis true.

In making this comparison, we must imagine two kinds of counter factuals in hypothesis testing. (1) What if p_0 is right? How would the data look? (2) What if p_0 is wrong and in stead some other argue is right which predicts p equal some other value, say p_A? How would the data look under various alternative theories and values of p?

This framework for testing hypotheses creates a dichotomy between the hypothesis and *not* the hypothesis. To construct a test of the hypothesis we think conditionally. If the hypothesis is true, what are the chances of observing the data that we have observed? If the hypothesis is untrue, what are the chances of observing the data? These possibilities are summarized in the table.

Hypothesis Framework

	Data Indicate Hypothesis is	
Hypothesis is	True	False
True	Correct	False -
False	False +	Correct

We employ the data in making two sorts of probability calculations.

First, what is the probability of observing the data if the hypothesis is true? The hypothesis in our simple example implies that the true proportion is p_0 and that the variance of X is $p_0(1-p_0)$. Hence, we want to calculate:

$$P(|\hat{p} - p_0)| > z_0\sqrt{p_0(1-p_0)/n}|p = p_0) \leq \alpha.$$

We can use the normal probability to make this calculation. Assuming $\alpha = .05$, $Z_0 = 1.964$. We, then, calculate whether the statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

exceeds 1.96. If the estimated p deviates sufficiently from p_0, then we conclude that the data do not support the hypothesis, because the data were unlikely to have occurred by chance were the hypothesis true.

In the example of the LA County data,

$$P(|z| > z_0) = P(|z| > \frac{.125 - .1}{\sqrt{\frac{.1.9}{2500}}}) = P(|z| > 4.16) < P(|z| > 1.96) = .05$$

So, the data do not support the working hypothesis of .1.

Notice that the hypothesis test that we implemented was extremely similar to the confidence interval calculation. Indeed, there is a duality between these ideas. An equivalent way to conduct hypothesis tests if to ask whether the hypothesized value falls inside the 95 percent confidence interval or not. One caveat is that the confidence interval calculation must use information about the variance as well as the mean under the null hypothesis. The appropriate calculation of the confidence interval for the hypothesis test is $\hat{p} + / - 1.96\sqrt{p_0(1 - p_0)/n}$.

This is a subtle difference that usually doesn't matter in practice. It does reflect the fact that you are conditioning on a hypothesis and all that implies about the distribution of the data.

We have ignored the other conditional, which is sometimes referred to as power. This calculation involves entertaining alternative hypotheses and performing similar calculations to the ones above. Power is useful in designing studies. Specifically, power amounts to asking how much ability does your study have to distinguish hypotheses. This depends on the amount of information you have collected.

1.D. Design

Design of studies involves a large number of choices. What are the key variables? How are they to be measured? How are the data to be collected (e.g., sample frames)? How many cases must be observed? All of these issues are important in study design. If, for example, we have a lot of measurement error, then the confidence intervals will be inflated. If we have bias then the confidence intervals will be wrong.

A basic design choice is sample size. In order to be able to distinguish among alternatives how much data do I have to have? Rather than develop the idea of power fully, I will show you a basic short cut.

Before doing a study, we must decide how much we desire to be able to discriminate across possible values of the parameters. In a survey, for example, we might choose to estimate a proportion within, say, 3 percentage points. Call this level of confidence L. We wish, then, to be able to make a statement such as "I'm 95 percent sure that the true value lies in $\hat{p} + / - L$." We call L the margin of error; we choose a value for L.

Once we collect the data we know how we will analyze it. We will construct a 95 percent confidence interval using the normal probability approximation and the sample estimates. That is, $\hat{p} + 1.96\sqrt{p(1-p)/n}$.

A handy formula for computing sample sizes emerges when we compare these two simple formulas. One formula expresses what we wish to be able to say; the other expresses what we will be able to say. The first term in each formula is the same. To square are wishes with our abilities, let $L = 1.96\sqrt{p(1-p)/n}$. Solve for n:

$$n = \left(\frac{1.96}{L}\right)^2 p(1-p)$$

To calculate this value we need only make a guess about p. The most conservative guess is $p = .5$. Let $L = .03$ – a commonly used margin of error for proportions. Hence,

$$n = \left(\frac{1.96}{.03}\right)^2 (.5)(.5) = 1067$$

To have a relatively tight margin of error around a sample proportion, one needs to sample at least 1000 people.

One general lesson about design from this calculation is that the design of studies consists of "thinking backward." In designing a study, think about how the data are to be analyzed and what hypothesis tests are to be conducted. This will guide decisions about sample size and measurement.

2. Central Limit Theorem

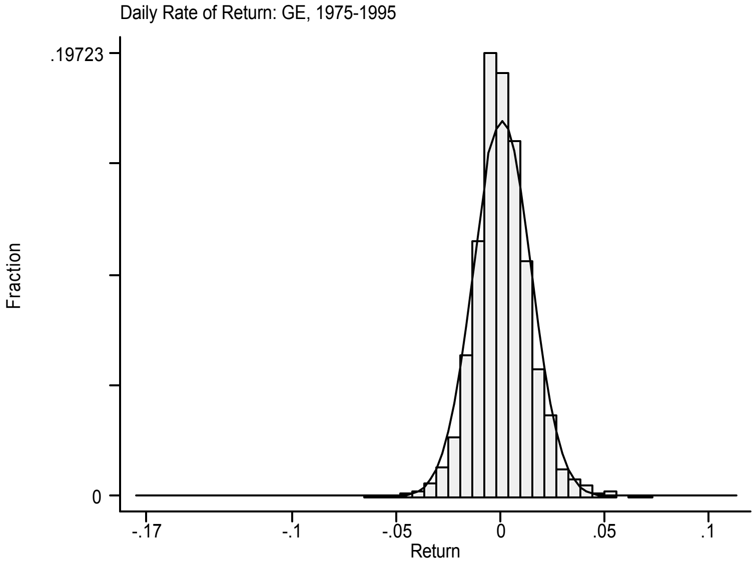
So far, we have developed our methods of inference and estimation case-by-case. A specific problem has a particular distribution, which leads to a specific estimator and inferential distribution. A very powerful and elegant theorem unifies statistical methods, and that is called the Central Limit Theorem. The Central Limit Theorem states that the sum of random variables is itself a random variable and follows a normal distribution, with mean $n\mu$ and variance $n\sigma^2$. Because most data consist of sums of random variables, the normal distribution is a starting point for statistical modeling. And, any inferences we wish to draw about means, regression lines, and other quantities of interest are made based on the normal distribution.

To underscore the idea that normality approximates behavior well, consider two important substantive examples.

Example 1. Stock Markets.

The graph shows the distribution of the daily rate of return on General Electric Stock from 1975 to 1995. The rate of return is the percent change in the stock's value. The graph shows the returns for a span of over 5000 days. General Electric's average rate of return in the 20 years is .0008 – just under one-tenth of one-percent per day. The variance of this stock (sometimes taken as a measure of risk) is .0002. Overlaid on the histogram of daily rates of return is the normal curve with a mean of .0008 and variance .0002.

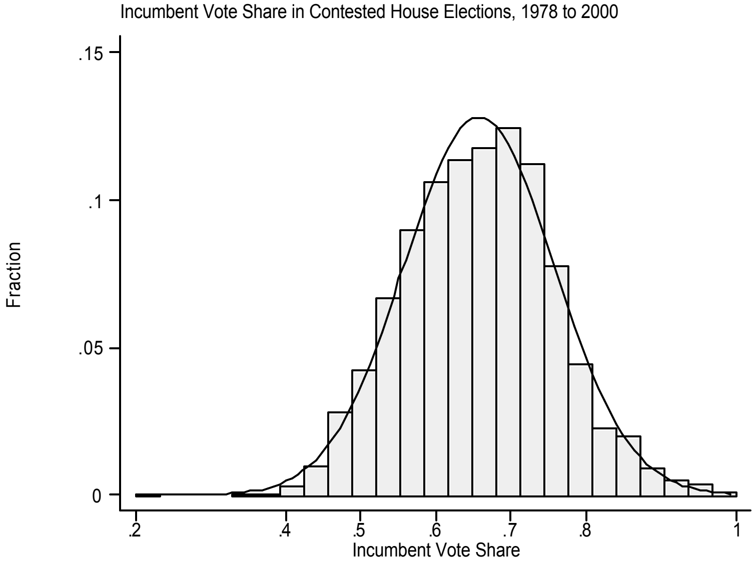
Two features of GE's rate of return deserve note. First, the distribution looks very normal – a symmetric bell shaped curve and a strong central tendency. Normality, then, may be a very good approximation. Second, in the details, the data deviate from normality in some interesting ways. There is a large negative outlier at -.17, corresponding to the crash of 1987, and a large positive outlier at .11, corresponding to a correction to the crash. The data look too Kurtotic. The mean is .0008, the variance .0002, the skew is approximately 0, but the kurtosis (fourth moment from the mean) is 11. With the normal distribution one expects a kurtosis around 3. This says that there are too many extreme deviations. If one is thinking about markets generally, the intuitions from the normal may be quite good for an approximation. If one is trying to model day to day behavior and make money on large volumes of trading, deviations from normality may be quite important.



Example 2. Elections.

F.Y. Edgeworth observed in the Journal of the Royal Statistical Society in 1898 that the distribution of the vote should be normal. He reasoned that the fraction of pro-Conservative voters (in England) in the population is p and that a "sample" of n people vote. Hence, the distribution of the vote in a constituency is Binomial, which is approximated well by the normal distribution. Kendall and Stuart (British Journal of Sociology 1951) developed this thinking into a model of the distribution of votes across districts, and the normal distribution has since become the standard model for thinking about the variation in votes across districts and over time.

The graph shows the distribution of the incumbent candidate's share of the two-party vote in U.S. House elections from 1978 to 2000. The variable equals the Democrat's share of the vote when the incumbent is a Democrat and the Republican's share of the vote when the incumbent is a Republican. There are 3630 district-level election outcomes in the data. The average incumbent vote share is .66 and the standard deviation is .10. The normal curve with mean .66 and variance .01 is overlaid on the histogram. As with the stock data the distribution of votes across districts is approximated well by the normal distribution. The curve deviates from normality somewhat, showing a slight positive skew.



Stocks and votes are examples of normality approximating behavior that is the sum of many smaller actions or events. Statistics, such as the sample mean and sample variance, are similarly sums of random variables. Hence, the normal distribution and the Central Limit Theorem unify and simplify statistical analysis and inference.

We will derive a version of the Central Limit Theorem for the sum of n Bernoulli random variables, though a general proof of the Central Limit Theorem is beyond the scope of this course. Before presenting analytical results, we develop the intuition behind the Central Limit Theorem using simulations.

2. A. Simulations

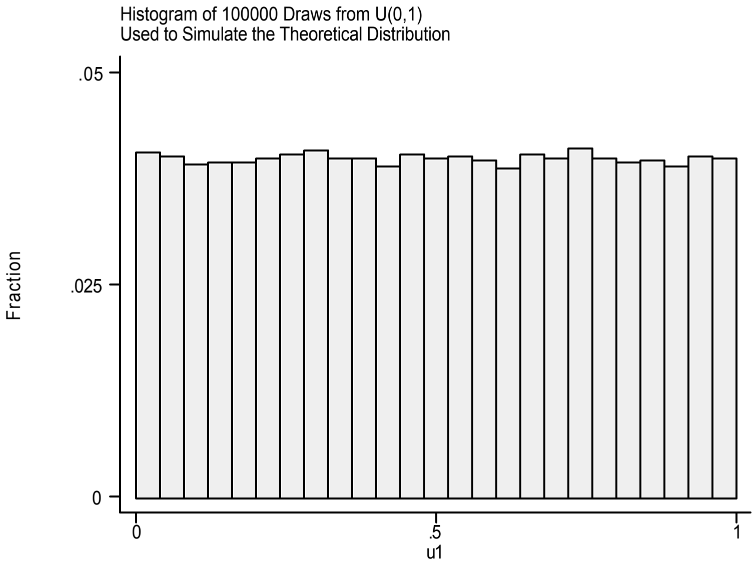
The goal of these simulations is to demonstrate that the sum of random variables each of which has very non-normal distribution tends to normality.

The simulations proceed in several steps. First, we generate a large number of draws (100,000 in the first graph) from a uniform distribution to approximate the uniform density function. I am using the uniform as an example of a non-normal distribution. Second, we simulate many such uniform random variables. Each simulated distribution is the density function of an independent random variable. Third, we consider progressively larger sums of uniform densities to see how quickly the normal density emerges.

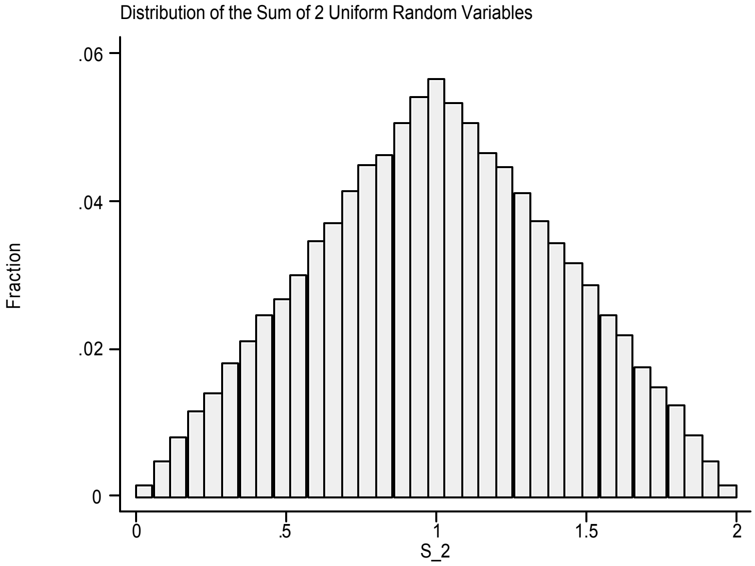
We will do two such simulations. The first corresponds to sums of uniform random variables. One could do this problem analytically using convolutions, as in Bulmer Problem 3.5. The distribution of the sum of two uniforms is Triangular, etc. Here we will let the computer do the math for us. The second corresponds to the distribution of statistics. If we take n draws from the uniform as mimicking the behavior of sampling from a population, we can study the distribution of repeated samples from the distribution. What is the distribution of the mean? What is the distribution of the standard deviation? How does the sample size affect the distribution?

Example 1. Sums of Uniform Random variables

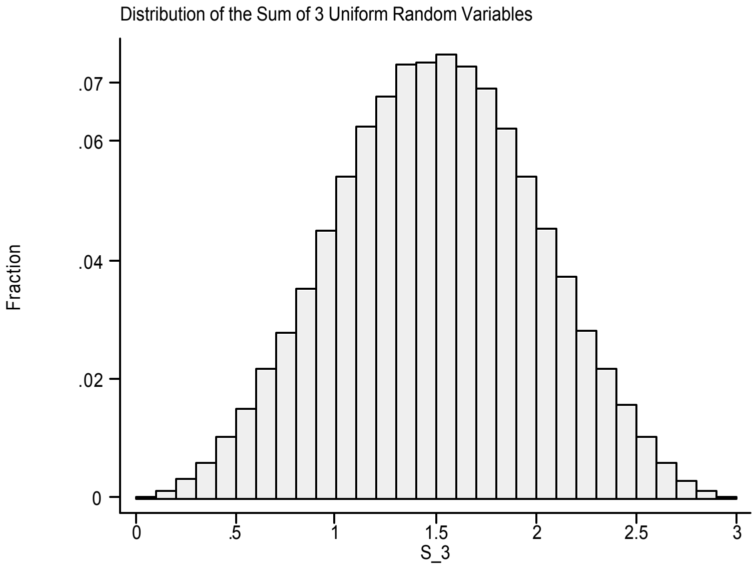
The uniform distribution ranges from 0 to 1, has a mean of .5 and variance of $\frac{1}{12}$. From the properties of expected values we can make some basic conjectures about the distribution of the sum of uniform random variables. The sum of two uniform distributions has a minimum value of 0 and a maximum of 2; it will have a mean of $1 = 2(.5)$ and a variance of $\frac{2}{12}$. The sum of n uniform distributions has a minimum value of 0 and a maximum of n; it will have a mean of $\frac{n}{2}$ and a variance of $\frac{n}{12}$.



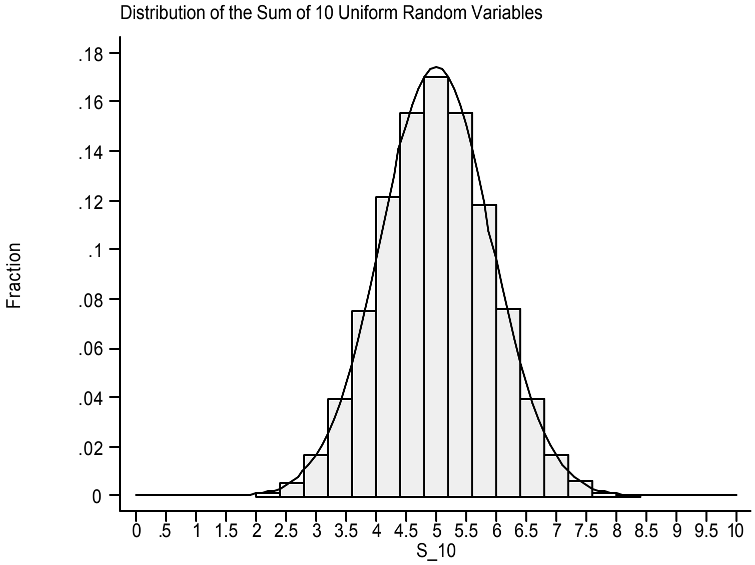
Histogram of 100000 Draws from U(0,1)
Used to Simulate the Theoretical Distribution



Distribution of the Sum of 2 Uniform Random Variables



Distribution of the Sum of 3 Uniform Random Variables



Distribution of the Sum of 10 Uniform Random Variables

The first graph shows the density of the random variable. The second graph shows the result from the problem set that the sum of two uniform random variables has a triangular density. Summing 10 uniforms produces a highly uniform density function.

To build your intuition about this pattern, convert the continuous problem into a discrete problem. Divide the Uniform into two equal parts; let the first part have value 0 and the second part have value 1. If we add two uniforms together we have 3 possible values – 0, 1 and 2. The probability associated with any particular combination of 0's and 1's from variable 1 and variable 2 is .25. That is, the probability that we drew a value from the interval of the first uniform associated with 0 is .5 and the probability that we drew a value from the interval of the first uniform associated with 0 is .5. Hence, the probability of a $0, 0$ is .25; the same is true for $0, 1$, $1, 0$, and $1, 1$. There is one way to get a sum of 0; there are 2 ways to get a sum of 1; and there is one way to get a sum of 2. If we divided the uniform interval into, say, 10 subintervals and gave those intervals value 0 to 9. There would be one way to get a sum of 0, two ways to get a sum of 1, four ways to get a sum of 4, etc. We can make the subintervals as small as desired and arrive at a continuous triangular distribution for the sum of two random variables.

The mathematical structure behind this result is called a convolution. Let $U = X + Y$. We want to derive the distribution $H(u) = P(X + Y \leq u)$. We can derive this from the conditional distribution of $X|Y$ and the distribution of X. Let F be the distribution of $X|Y$ and G be the distribution of y. For any given u and y, we have $P(X \leq u - y) = F(u - y|Y = y)$, so $P(U \leq u) = \sum_y F(u - y)g(y)$ if y and x take discrete values and $P(U \leq u) = \int F(u - y)g(y)dy$.

Using the same reasoning we could add as many uniform random variables together as we wished. The intuition for what is going on derives from the binomial coefficients. What drives the central limit theorem is the number of ways that one can get specific values for the sum, which is determined by the binomial coefficients. Dividing the uniform into 2 parts and adding n uniforms together will produce a binomial distribution, which is approximated very well by the normal.

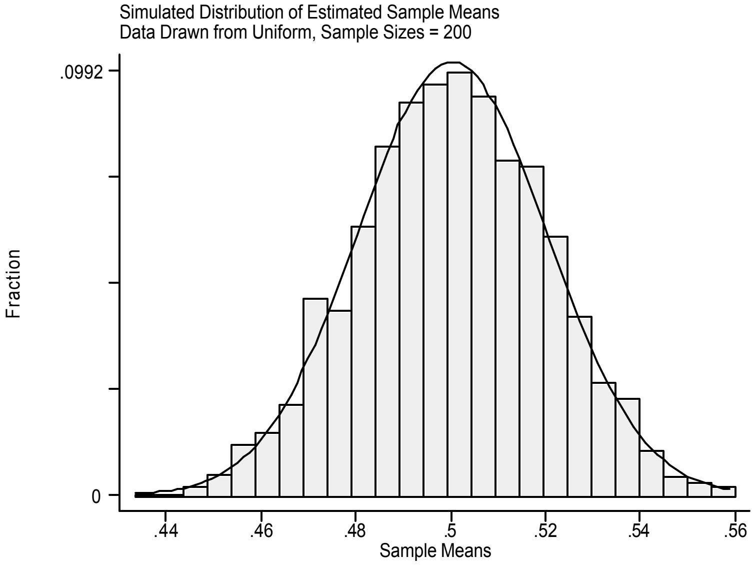
Example 2. Distributions of Sample Statistics

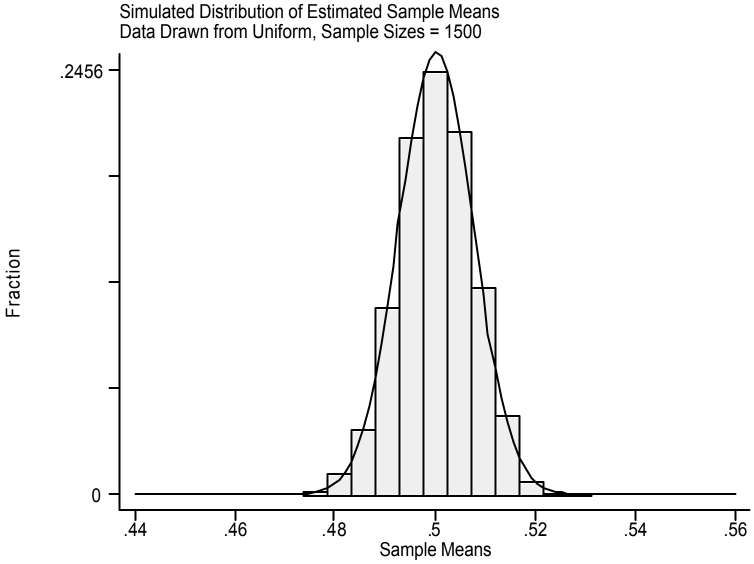
Suppose that we took a sample of size n from the population. Regardless of the population distribution, the distributions of the sample statistics will be approximately normal if n is large enough. We demonstrate that here by considering two cases of random draws from the uniform distribution. Expected value of any draw from the uniform distrution is .5 and the variance $.288 = \sqrt{1/12}$. The sample statistics $\bar{x}$ and $\hat{\sigma}^2$ should, then, be close to .5 and .288.

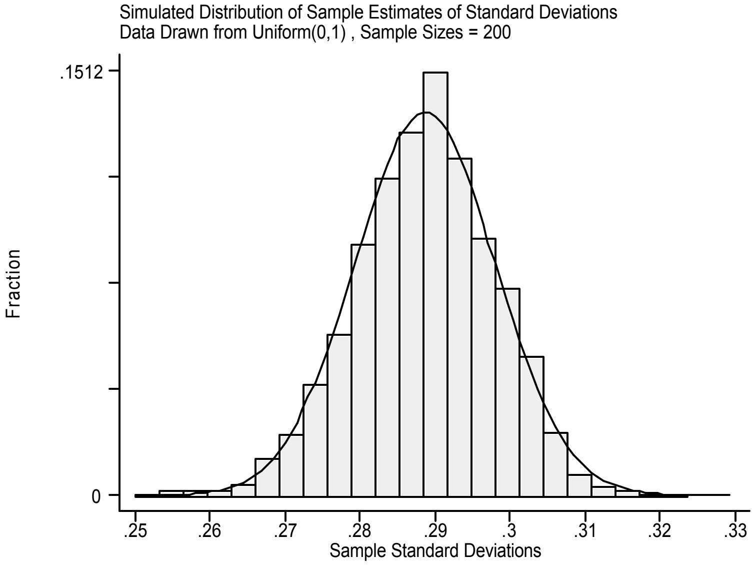
How close depends on the sample size. We know from the theoretical results derived earlier that the expected value of the sample average is the true mean and the variance of the sample average is σ^2/n. I performed two simulations. One models a sample of 200 from a uniformly distributed random variable; the other models a sample of 1500 from a uniformly distributed random variable. To simulate the distribution of the estimated mean and variance I drew 2500 such samples and mapped the histogram of the means and variances using the following STATA code:

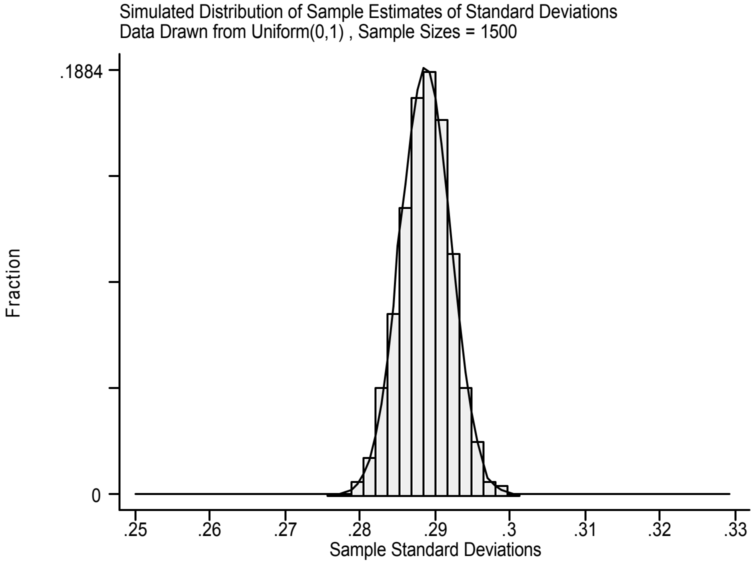
```
set mem 300m
set obs 1500
forvalues i = 1(1)2500 f
      gen u_‘i’ = uniform()
      g
forvalues i = 1(1)2500 f
      quietly sum u_‘i’
      disp r(mean) r(sd)
      g
```

The distributions show two striking patterns. First, the distributions of the means and variances are approximated very well by the normal distribution. Second, the distribution of the sample means and variances is much tighter in the larger sample. This is an example of the law of large numbers. From theory, we expect that the variance of the mean is $\frac{1/12}{n}$, so the distribution of sampling means shrinks at rate $1/\sqrt{n}$. The sample of 200 should have standard error roughly 2.7 times larger than the sample of 1500.


Simulated Distribution of Estimated Sample Means
Data Drawn from Uniform, Sample Sizes = 200
.0992
0
Fraction
.44
.46
.48
.5
.52
.54
.56
Sample Means


Simulated Distribution of Estimated Sample Means
Data Drawn from Uniform, Sample Sizes = 1500
.2456
0
Fraction
.44
.46
.48
.5
.52
.54
.56
Sample Means





Some statistics are not sums – for example, the median. Many statistics used in lieu of sums (such as the median instead of the mean) rely on orderings of data. Inference with such statistics usually depends on the population or sampling distribution of the random variable, complicating statistical inference. Sometimes medians and mean absolute deviations are preferrable, as they guard against outliers. However, statistical analysis with statistics based on orderings is much more cumbersome. The core of statistical methods relies on averaging and the Central Limit Theorem simplifies statistics enormously.

2.B. Formal Proof

Here we derive a simple version of the Central Limit Theorem to strengthen your understanding of how the normal distribution approximates sums of random variables.

DeMoivre-Laplace Central Limit Theorem. Let S_n be the sum of n independent and identical Bernoulli trials with probability p. For some numbers, z_1 and z_2, as n gets large

$$P(np - z_1\sqrt{np(1-p)} \leq S_n \leq np + z_2\sqrt{np(1-p)}) \rightarrow \Phi(z_2) - \Phi(z_1),$$

where $\Phi(\cdot)$ is the cumulative distribution of the standard normal distribution.

The proof of this result proceeds in two steps. First, we consider the central part of the distribution, which is the probability in the vicinity of the mean $m = np$. We will show that the density at the mean is $a_0 = \frac{1}{\sqrt{2\pi np(1-p)}}$. This result follows immediately from Stirling's formula and is the "central limit." Second, we consider the density associated with specific deviations around the mean, indexed by k. We will show that the density a_k at the point k is approximately $a_0 e^{-\frac{1}{2np(1-p)}k^2}$. This result follows from the approximation of the ratio of two series of numbers.

Proof (due to Feller). The sum of n independent Bernoulli random variables follows the Binomial distribution. Define the mean of the of S_n as $m = np$. We will assume for simplicity that n is such that m is an integer. If not, we would add to this quantity a fractional amount δ to make m the nearest integer value. Because this component is inessential to the proof we will ignore this term.

First, analyze the central term of the binomial, i.e. $S_n = m$. Let,

$$a_0 = \binom{n}{m} p^m (1-p)^{n-m} = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}.$$

From Stirling's Formula we know that we can approximate this function as

$$a_0 = \frac{\sqrt{2\pi n} n^n e^{-n}}{\sqrt{2\pi m} m^m e^{-m} \sqrt{2\pi (n-m)} (n-m)^{n-m} e^{-(n-m)}} p^m (1-p)^{n-m}$$

This reduces to

$$a_0 = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{m(n-m)}} \frac{n^m n^{n-m}}{m^m (n-m)^{n-m}} p^m (1-p)^{n-m}$$

Substituting the definition of $m = np$, we find that

$$a_0 = \frac{1}{\sqrt{2\pi}\sqrt{np(1-p)}}$$

Second, consider deviations around the central term. Let x be a negative or positive integer such that for all values of S_n,

$$a_x = \binom{n}{m+x} p^{m+x} (1-p)^{n-m-x}$$

This is the formula for the binomial for all values from 0 to n, where I have rewritten the index value such that x ranges from $-m$ to $n-m$, instead of the usual indexing of k ranges from 0 to n.

We wish to express the density at any point as a function of the deviation from the central tendency. Notice that the binomial above has in it the expression for the density at m, i.e., a_0. Consider the ratio

$$\frac{a_x}{a_0} = \frac{\frac{n!}{(m+x)!(n-m-x)!}}{\frac{n!}{(m)!(n-m)!}} \frac{p^{m+x}(1-p)^{n-m-x}}{p^m (1-p)^{n-m}}$$

This reduces to

$$\frac{a_x}{a_0} = \frac{(n-m)(n-m-1)...(n-m-x+1)p^x}{(m+1)(m+2)...(m+x)(1-p)^x}$$

We can rewrite this term as

$$\frac{a_x}{a_0} = \frac{(1 - pt_0)(1 - pt_1)...(1 - pt_{x-1})}{(1 + (1 - p)t_0)(1 + (1 - p)t_1)...(1 + (1 - p)t_{x-1})},$$

where $t_j = \frac{j+(1-p)}{np(1-p)}$. [Note: Verify that this substitution works.]

Finally, we can analyze the ratio on the right-hand side of this expression using Taylor's expansion for the natural logarithm.

$$log(1 + (1 - p)t_j) = (1 - p)t_j - \frac{1}{2}[(1 - p)t_j]^2 + \frac{1}{3}[(1 - p)t_j]^3 - \frac{1}{4}[(1 - p)t_j]^4...$$

and

$$log(\frac{1}{1 - pt_j}) = (p)t_j + \frac{1}{2}(pt_j)^2 + \frac{1}{3}(pt_j)^3 + \frac{1}{4}(pt_j)^4...$$

Adding these two terms we get:

$$log\left(\frac{1 + (1 - p)t_j}{1 - pt_j}\right) = t_j - \frac{1}{2}t_j^2(1 - 2p) + \frac{1}{3}(t_j)^3(p^3 + (1 - p)^3)...$$

The terms above t_j are small because they are multiplied by fractions p. Hence, $log(\frac{1+(1-p)t_i}{1-pt_j}) \approx t_j$ or

$$\frac{1 - pt_j}{1 + (1 - p)t_j} \approx e^{-t_j}.$$

The last expression captures the deviation for just one term in the expansion for a_k/a_0. Hence,

$$a_k = a_0e^{-(t_0+t_1+...t_{x-1})}$$

Because the sum of the first x numbers equals $x(x + 1)/2$, we can use the definition of t_j to rewrite the exponent as

$$t_0 + t_1 + ...t_{x-1} = \frac{0 + 1 + 2 + ...(x - 1) + x(1 - p)}{np(1 - p)} = \frac{x(x - 1)/2 + x(1 - p)}{np(1 - p)} \approx \frac{x^2/2}{np(1 - p)}$$

Pulling all of the pieces together reveals that

$$a_x = a_0e^{\frac{x^2/2}{np(1-p)}} = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{np(1 - p)}}e^{\frac{x^2/2}{np(1-p)}}$$

, which is the formula for the normal density. This proves that for the sum of n independent Bernoulli trials, any deviation x away from the mean of the sum, $m = np$, is approximated by the normal distribution with variance $np(1-p)$.

A more general proof of the Central Limit Theorem, for random variables that follow other (almost any) distributions, is offered in Bulmer and relies on moment generating functions. The power of the Central Limit Theorem is that for nearly any distribution, the behavior that consists of sums of random variables (such as votes and stock markets) follows a normal distribution and the distribution of averages and other sample statistics based on sums will follow the normal distribution.

3. Means and Variances

The frequency of any random variable can be studied using the population (or theoretical) mean and variance. The Bernoulli is a special case of this where the parameter of the distribution function is p and $E(X) = p$ and $V(X) = p(1-p)$. Typically, we will have a random variable that takes more complicated values and frequency functions. Our task is to estimate the mean and variance from a sample of data and perform inference based on that information.

Voting provides an instructive examples of the varieties of distributions that commonly arise. We might wish to study the frequency of victory – the probability that candidates of a certain type of party or from a particular social group win. Grofman, for instance, studies the probability of reelection of incumbents. Political scientists also study the aggregate of the election outcomes – the distribution of seats or division within Congress or the parliament. Finally, since winning depends on the vote, we can often learn more about the forces affecting elections by studying the behavior of the vote. What is the distribution of the vote across districts? What is the variability of the vote within a typical district or a type of district over time?

Means, variances, and histograms (density estimates) are staples in the study of elections. Some theories also lead us to focus on other quantities. Most important of these is the median. An important theoretical tradition holds that parties and candidates contest elections by announcing policies in order to appeal to the greatest fraction of the electorate. In two party systems, competition for votes drives the parties to locate at the ideal policy of the median voter (Hotelling 1927). Similarly, in the study of legislative politics the median voter along a policy dimension is pivotal (Black 1968, Krehbiel 1998). In some problems, then, we wish to study the median as a theoretical matter. In general, means will be more efficient and, thus, a preferred statistic.

To estimate the mean and variance of a distribution, we begin with a summary of the data and proceed to choose values that satisfy one of our estimation criteria – such as the method of moments or maximum likelihood. I will present the maximum likelihood estimates here.

Data consist of n observations from a Normal distribution with mean μ and variance σ^2. This is often written $X_i \approx N(\mu, \sigma^2)$. Assuming the observations are independent, the likelihood function for these data is:

$$L(\mu, \sigma^2) = \Pi_i^n f(x_i) = \Pi_i^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_i^n (x_i-\mu)^2}$$

And the log-likelihood function is

$$ln(L) = \frac{-n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_i^n (x_i - \mu)^2$$

To derive the maximum likelihood estimators set the partial derivatives of $ln(L)$ with respect to μ and σ^2 equal to 0.

$$\frac{\partial ln(L)}{\partial \mu} = -\frac{1}{2\hat{\sigma}^2} \sum_i^n (-2)(x_i - \hat{\mu}) = 0$$

$$\frac{\partial ln(L)}{\partial \sigma^2} = \frac{-n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \sum_i^n (x_i - \hat{\mu})^2 = 0$$

Solving the first equation reveals that

$$\hat{\mu} = \bar{x}$$

Substituting this result into the second equation and solving yields:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i^n (x_i - \bar{x})^2$$

These are the sample averages and variances.

The properties of these estimators can be derived from their means and variances. We know that if the sample is large the distributions of these two statistics will be approximated by the Normal distribution.

We have already derived the mean and variance of the sample average.

$$E(\bar{x}) = \mu$$

$$V(\bar{x}) = \frac{\sigma^2}{n}$$

When we must estimate σ, the estimated variance of $\bar{x}$ is $\frac{\hat{\sigma}^2}{n}$

Similarly, we can estimate confidence intervals using the approach outlined above. Consider the Incumbency data. From 3630 US House elections from 1978 to 2000, the average incumbent vote margin is .66 with an estimated standard deviation is .10. Hence, the standard error of the mean incumbent vote margin is $.0017 = .1/\sqrt{3630}$. A 95 percent confidence interval for the mean incumbent vote margin is [.657, .663].

As another example, consider politicians' ideologies. A group administered the National Political Aptitude Test to candidates for Congress in 1996 and 1998. Ansolabehere, Snyder and Stewart (2000) construct a preference score. The distribution of that score is shown for incumbents. The mean score is .54 and the variance is .28 for the 376 incumbents in the sample. A 95 percent confidence interval for the true mean is $.54 \pm 1.96\sqrt{.28/376} = .54 \pm .053$. This is the confidence interval for an estimate of the center of the distribution. It is not a necessarily a good prediction of what score an individual incumbent might receive. To do that more information about the district and race is important.

By way of contrast, consider the distribution of the median, x_r, where r is the 50th percentile case. Generally, this estimator is biased for the mean, μ: $E[x_r] = \mu$ if $f(x)$ is symmetric. Assuming that it is biased, the sample median is usually inefficient. If $f(x)$ is symmetrical around μ, it can be shown that the lower bound of the variance of the median is:

$$Var(x_r) \geq \frac{1}{4[f(\mu)]^2(n+2)}$$

To see that this usually exceeds the variance of the mean consider the uniform and normal cases. When $f(X)$ is uniform on the interval $(0,1)$, $V(x_r) = \frac{1}{4(n+2)}$. The variance of the sample average from n observations is $\frac{\sigma^2}{n} = \frac{1}{12n}$. In this example, the sample average is approximately 3 times more efficient than the sample median as an estimate of the mean of the distribution. When $f(X)$ is normal with mean μ and variance σ^2, the lower bound for $V(x_r) = \frac{2\pi\sigma^2}{4(n+2)}$. This is approximately 1.6 times larger than $\frac{\sigma^2}{n}$.

As mentioned, there are some problems where we are interested in the median of the distribution, rather than the mean. In the study of legislative committees the median voter is pivotal. What is the median of the 25 U.S. House committee members? Roll call voting studies average sets of votes and estimate the ideal point of each legislator on a continuum of left-to-right policy preferences. Researchers would like to know where the median of the committee lies. One way to estimate a 95 percent confidence interval for the median is to use the normal approximation above. The resulting interval is approximately, $x_r \pm 1.96\sqrt{1.57\frac{\sigma^2}{4(n+2)}}$.

Consider the case of Incumbents' ideologies. The median score is .59, and the 95 percent confidence interval (using the normal approximation) is $.59 \pm 1.96\sqrt{2\pi\frac{\sigma^2}{4(n+2)}} = .59 \pm (1.96)(.034) = .59 \pm .067$

The estimated variance also has a sampling distribution. We consider it's properties briefly here.

The estimate of the variance is biased, but consistent.

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n}(x_i - \mu + \mu - \bar{x})^2\right]$$

Expanding the square and collecting terms yields

$$E[\hat{\sigma}^2] = \frac{1}{n}E\left[\sum_{i=1}^{n}(x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) + (\mu - \bar{x})^2\right] = \frac{1}{n}\sum_{i}\left[\sigma^2 - 2E[(x_i - \mu)(\bar{x} - \mu)] + E[(\mu -\right.$$

The middle term in the last part of the equation equals $-2E[(x_i - \mu)(\frac{1}{n}\sum_{j=1}^{n}(x_j - \mu)] = -2\frac{\sigma^2}{n}$, because $Cov(X_i, X_j) = \sigma^2$ if $i = j$ and 0 otherwise. The last term in the equation above equals $\frac{1}{n}\sigma^2$, because this is just the variance of the mean. Hence,

$$E[\hat{\sigma}^2] = \frac{1}{n}\sum_{i}[\sigma^2 - \frac{\sigma^2}{n}] = \sigma^2 - \frac{1}{n}\sigma^2$$

This does not equal σ^2, so the estimtor is biased. It is consistent because the bias is of the order $1/n$.

Bias in the variance is corrected with the appropriate "degrees of freedom." An alternative, unbiased estimator is usually employed in computations:

$$s^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$$

The "degrees of freedom" refers to the fact that we have n pieces of information and we "used" one of them to estimate the sample mean. If we had only one observation, there would be no information about the variance.

Finally, note that the estimated variance follows a χ^2 distribution with $n-1$ degrees of freedom. The estimated variance is the sum on $n-1$ independent normal variables. To derive this distribution, assume that the data are approximately normal. Standardize each observation and square it. This gives $\frac{(X_i - \bar{X})^2}{\sigma^2}$. Each of these is standard χ^2 random variable. Summing yields $\frac{\sum_i (X_i - \bar{X})^2}{\sigma^2}$, also a χ^2 distribution, but with $n-1$ degrees of freedom. This distribution has a mean of $n-1$ and a variance of $2(n-1)$.

Distributions of variances are very important in statistics, perhaps as important as distributions of means. First, sometimes we wish to make tests about variances directly. Second, in confidence intervals for means we must adjust for the distribution of the variance if we had to estimate the variance in calculating the standard error. Third, hypothesis tests may be thought of as the distance between our observations and our expectations. We can formalize this into test statistics that take the form of variances – squared distances between hypothesized values and observed values.

Here I will not go into tests about variances, as they are sufficiently rare. Instead, let us turn to the problem of confidence intervals when we must estimate the variance.

In smaller samples, we must worry about an additional issue – the fact that we estimated σ^2 in constructing the confidence interval. The confidence interval captures the probability that a deviation from the mean lies within a set interval. That is, a $1-\alpha$ confidence interval is

$$P(-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \leq z_{\alpha/2}) = P(\bar{x} - z_{\alpha/2}\sigma_{\bar{x}} \leq \mu \leq \bar{x} - z_{\alpha/2}\sigma_{\bar{x}})$$

We can use the normal because $\bar{x}$ (or some other estimator) is a sum of random variables.

However, if $\sigma_{\bar{x}}$ is estimated, it too is a random variable and we must account for this fact.

The inner term in the confidence interval is now $\frac{\bar{x}-\mu}{\hat{\sigma}/n}$. This function is a random variable that is the ratio of a normally distributed variable to the square root of a Chi-squared distributed variable. Bulmer offers a derivation of the resulting distribution, which is Student's T-distribution.[1] The T is symmetric, like the normal, but slightly thicker in the tails. The values of the T-distribution that cover $1-\alpha$ percent of the distribution depend on the degrees of freedom of the estimated standard deviation. We will use the notation $t_{\alpha/2,n-1}$ to denote the relevant cut points. The confidence interval can now be expressed as:

$$P(-t_{\alpha/2,n-1} \leq \frac{\bar{x}-\mu}{\hat{\sigma}/n} \leq t_{\alpha/2,n-1}) = P(\bar{x} - t_{\alpha/2,n-1}\hat{\sigma}/n \leq \mu \leq \bar{x} - t_{\alpha/2,n-1}\hat{\sigma}/n)$$

For $\alpha = .05$, when $n = 10, t_{\alpha/2,n-1} = 2.23$; $n = 20, t_{\alpha/2,n-1} = 2.09$; when $n = 50, t_{\alpha/2,n-1} = 2.01$; when $n = 75, t_{\alpha/2,n-1} = 1.99$; when $n > 120, t_{\alpha/2,n-1} = 1.96.$. Rule of thumb, use $t > 2$.

Wald generalized the idea of the t-statistic and t-test into an overall approach to statistical testing. Wald tests consist of calculating the distance between the observed data and an hypothesized outcome, normalizing using the variability in the data implied by the hypothesis. The general form of the Wald test is the square of a T-statistic. This ratio is the square of a normal distribution divided by the square of another normal distribution. This is the F-statistic. The F-statistic is the ratio of two "variances". In the numerator of the Wald test is the mean squared error under a null hypothesis (the distance between the estimated parameter and the hypothesized). In the denominator is the sampling variance, which is the distance one would expect if the deviations were not systematic, and just due to sampling variance.

A special case is Pearson's Chi-squared test for independence of two random variables in a table. Assume there are two variables X and Y, which have C (for columns) and R (for rows) values respectively. The table that characterizes the joint distribution has $C \times R$ cells. The count in each cell is a binomial distribution, where the probability of being in the cell is the probability of the pair $(X = c, Y = r)$, i.e., $P_{c,r}$. The expected count in each

[1] The T-distribution was derived by Gossett, a quality control engineer and statistician employed by Guiness. He was not allowed to publish under his own name and instead used the nom de plume Student.

cell is $nP_{c,r}$ and the observed count is $k_{c,r}$. For convenience suppose that we relabel the indexes so that $m = 1, 2, ..., CR$. For example in a 2 by 2 table, we may relabel the indexes so that $(X = 0, Y = 0)$ is $j = 1$, $(X = 1, Y = 0)$ is $j = 2$, $(X = 0, Y = 1)$ is $j = 3$, and $(X = 1, Y = 1)$ is $j = 4$. [It really doesn't matter how we relabel the index so long as we are consistent.]

Consider the null hypothesis that the two variables are independent. The expected count in each cell equals the number of cases total times the probability of observing an observation in that cell. Under the null hypothesis of indepedence, $P(X = c, Y = r) = P(X = c)P(Y = r)$. Hence, if the null hypothesis is true, the expected cell counts are $E_j = nP(X = c)P(Y = r)$. The squared deviation of each cell count from its expected count follows a χ^2-distribution. We must normalize by each cells variance, which is approximately equal to $nP(X = c)P(Y = r) = E_j$. Pearson's Chi-squared statistic, then, is

$$\sum_j \frac{(k_j - E_j)^2}{E_j}.$$

As the sum of squared normals this follows a χ^2-distribution. The numerator is the deviation of the observed from the expected and the denominator is approximately the variace expected if the null hypothesis is true.

How many independent variables are in the sum? That is, how many degrees of freedom are there? Answer: there are $(R - 1)(C - 1)$ degrees of freedom. We estimate all of the row and column (marginal) probabilities in order to estimate the expected outcome under the null hypothesis. Within any row, then, if we know any $C - 1$ numbers, then we know the Cth, because the rows sum to 1. Similarly within any column, if we know any $R - 1$ numbers, then we know the Rth, because the rows sum to 1. In the entire table, then, there are $(R - 1)(C - 1)$ "free" numbers. Once we fill in values for these, the remaining numbers are determined. Hence the term "degrees of freedom."

Example. Lake, "Powerful Pacifists," APSR (1993). Are Democracies More Likely than Autocracies to Win Wars? Lake classified all wars in the 20th Century according to the type of regime involved and whether the regime was on the winning side.

	Nature of Regime		
Success in War	Democratic	Autocratic	
Win	38	32	70
Lose	9	42	51
	47	74	121

Independence implies that $P(D,W) = \frac{70}{121}\frac{47}{121} = .225$, $P(A,W) = \frac{70}{121}\frac{74}{121} = .354$, $P(D,L) = \frac{51}{121}\frac{47}{121} = .164$, and $P(A,L) = \frac{51}{121}\frac{74}{121} = .259$. The expected counts in the cells are $k_{D,W} = 27.2$, $k_{A,W} = 42.8$, $k_{D,L} = 19.8$, and $k_{A,L} = 31.2$. Pearson's Chi-squared statistic measures whether the observed counts were unlikely to have occured just by chance. This is calculated as:

$$X^2 = \frac{(27.2-38)^2}{27.2} + \frac{(42.8-32)^2}{42.8} + \frac{(19.8-9)^2}{19.8} + \frac{(31.2-42)^2}{31.2} = 16.673.$$

This follows a χ^2_1. The probability of observing deviations at least this large from the expected values is .00004 if the null hypothesis is true.

Why 1 degree of freedom? Given the marginal probabilities, once we observe one joint value, the other values are determined.

The probability of observing a deviation of the data from the expected value is called a *p-value*. STATA will calculate the value of the cumulative probability function for you (without a table). For the Chi-Squared test we can use the function **chi2(n, x)**. The command **disp 1 - chi2(1, 16.673)** returns the value **.00004**, which is the probability of observing a value of the χ^2_1 distribution at least as large as 16.673.

4. Effects and Differences of Means

Most often in the social sciences we study the relationship between variables. Specifically, we wish to know what the effect of one variable is on another.

Consider two examples.

Example 1. Regime Type and Success in War Lake's analysis of outcomes of wars shows that the type of regime is related to success in war. The analysis so far says nothing about how successful. A simple idea is to measure the difference between the probabilities of success of Democratic regimes and of Autocratic regimes. The win rate of Democratic regimes in his study is .81 (=38/47); the win rate of Autocratic regimes is .43 (=32/74). There is a 38 percentage point difference between Democrat's and Autocrat's win rates. Also, Democrtic regimes are twice as likely to be on the winning side as Autocratic regimes.

Example 2. Incumbency and Election Results We have seen that for incubments this distribution is remarkably normal. Were we to take a different perspective on these data a rather uneven picture would emerge. The figure shows the distribution of the Democratic share of the vote. The distribution of the Democratic share of the vote across districts has an uneven and at times bimodal distribution. Mayhew (1971) noticed this effect and attributed it to the rising vote margins of US House incuments. Erikson (1971) identified and estimated the *incumbency effect* as the differential in vote between those holding office and those not holding office. Over the last 20 years, the average division of the vote when there is an incumbent is .66 in favor of the incumbent's party; the average vote margin when there is no incumbent is .54 in favor of the party that previously held the seat. The interpretation commonly given is that there is a 12 percentage point incumbency advantage in votes – this is what we expect an incumbent to receive above and beyond what his or her party would win in the election in that constituency.

An *effect* of one variable on another is defined to be the difference in the conditional means across values of the conditioning variable. Suppose X takes two values, 1 and 0. The effect of X on Y is written $\delta = E[Y|X = 1] - E[Y|X = 0]$. This is the difference between the means of two conditional distributions, i.e., $\delta = \mu_1 - \mu_0$.

The effect is itself a quantity we would like to estimate and about which we would like to draw statistical inferences. How large is the difference in means? If we were to change the value of X, say through a public policy or experiment, how much change in Y do we expect? When we conduct a study and observe a difference between two subsamples (or subpopulations), how likely is it that it arose simply by chance?

4.A. Estimation

As a practical matter it will be convenient to use notation for random variables that clearly distinguishes the effect from chance variation in our random variables. Specifically, any single random variable can be written as the mean of the random variable plus an error term with mean 0 and variance σ^2: $Y = \mu + \epsilon$.

Using this notation, we may summarize the conditional distributions and effects as follows. Suppose X takes two values, indexed $j = 0, 1$. The conditional distribution of Y given X can be represented as:

$$(Y|X = j) = \mu_j + \epsilon_j,$$

where ϵ_j has mean 0 and variance σ_j^2. An alternative approach is to use the density function, $f(Y|X = j)$.

We can estimate the effect using the methods developed for a single mean. The theoretical effect is $\delta = E[Y|X = 1] - E[Y|X = 0]$, a parameter we don't observe directly. Data for this problem consist of *two* samples. One sample is drawn from the distribution of Y for the subpopulation of cases among whom $X = 1$, and a second sample is drawn from the distribution of Y for the subpopulation among whom $X = 0$. In many studies, the conditioning is not specified in the design of the data. Instead, we observe data from the joint distribution $f(Y, X)$ and then condition on values of X. For both situations, the difference of means is referred to as the "two-sample" problem.

The estimated effect of X on Y is the difference between the means of the two subsamples. Let n_1 be the number of observations for which $X =$ and let n_0 be the number of observations for which $X = 0$. Let i index observations within groups j. Then the estimated effect is:

$$d = \bar{y_1} - \bar{y_0} = \frac{1}{n_1} \sum_{i \in (X=1)} X_{i,j} - \frac{1}{n_0} \sum_{i \in (X=0)} X_{i,j}$$

The variance of each subsample is estimated using the familiar formula applied to the subsample:

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i \in (X=1)} (x_{i,j} - \bar{x_j})^2$$

What are the properties of the estimated effect, d?

1. It is unbiased.

$$E[d] = E[\bar{y_1} - \bar{y_0}] = E[\bar{y_1}] - E[\bar{y_0}] = \mu_1 - \mu_0 = \delta$$

2. The variance of the estimated effect depend on the nature of the sampling. Generally, we will encounter problems in which the two subsamples are independent.

$$V[d] = V[\bar{y_1} - \bar{y_0}] = V[\bar{y_1}] + V[\bar{y_0}] + 2Cov[\bar{y_1}, \bar{y_0}] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$$

When we estimate the sample variance we calculate the estimated $V[d]$ as $\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}$.

Importantly, the standard error of the estimated effect is bounded by the size of the smaller group. If we begin with a study in which n_1 is 100 and n_0 is 100, then the standard error will be of the order $1/\sqrt{100}$. If we increase the sample size of n_1 to, say, 1000, the standard error of the estimated effect is still of the order $1/\sqrt{100}$. Specifically, as $n_1 \to \infty$, $V[d] \to \frac{\sigma_0^2}{100}$.

In order to make efficiency gains from larger sample sizes, the optimal approach is to increase the sample sizes in both groups at about the same rate (depending on the relative variances).

In some problems, the samples are not independent and a somewhat different variance result holds. For example, some studies employ paired samples. Each observation in a sample is paired with another observation on the basis of a common characteristic that we think affects Y. One member of the pair is assigned $X = 0$ and the other is assigned $X = 1$.

The sample, then, consists of n pairs. The estimated effect is the average difference between each pair:

$$d = \frac{1}{n}\sum_{i=1}^{n} y_{i,1} - y_{i,0},$$

which is identical to the differnce between the averages of the two subsamples. The gain to pairing comes in the variance.

$$V[d] = \frac{1}{n^2}\sum_{i=1}^{n} V[y_{i,1} - y_{i,0}] = \frac{1}{n^2}(\sum_i V[y_i, 1] + \sum_i V[y_{i,0}] - 2\sum_i Cov[y_{i,1}, y_{i,0}] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0} - 2\frac{\sigma_i}{n},$$

where σ_i is the covariance between $1, 0$ within i. There is a gain in efficiency from pairing (a design effect) due to the covariance within pairs. To calculate the estimated variance of the effect we need only calculate the average sum of squared deviations of each paired difference from the average paired difference.

As an example, consider the "sophomore surge" as an estimate of the incumbency advantage. Erikson examines the change in the vote within congressional districts. Specifically, he looks at all races where the incumbent in time 1 was not the incumbent in time 0.

3. The sampling distribution of d is normal because it is the sum of random variables. That is

$$d \approx N(\mu_1 + \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0})$$

4.B. Inference about Effects.

Inference about an effect proceeds along the same lines as inference about a single mean. We may construct confidence intervals to measure our uncertainty about the estimated population parameter. We may also test whether the observed data deviate for a specified (or hypothesized) value.

Using the last property, we can draw inferences about the effect δ.

A 95 percent confidence interval can be constructed using Chebychev's inequality.

$$P(|d - \delta| > k\sqrt{V[d]}) < .05$$

From the Central Limit Theorem, we know that the distribution of d will follow a normal distribution if n_j is large. Therefore, $k = z_{\alpha/2} = 1.96$. Hence, a 95 percent confidence interval for the estimated effect is:

$$d \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}}$$

The law of large numbers applies to this interval, so long as both sample sizes grow as n grows. Specifically, as n_1 and n_0 get large, d approaches δ.

As with a single mean, when we have to estimate $V[d]$ we may need to use the T-distribution to correct for our uncertainty about the estimated variances. Now the choice of the T-distribution depends on the size of the smaller group. If it is below 100, the T is appropriate.

Hypothesis testing for a difference of means parallels the treatment for a single mean. We begin with a statement about a specific value $\delta = \delta_0$. The most common hypothesis for estimated effects is $\delta = 0$; hence the term "null" hypothesis for no effect.

If the hypothesis is true, then there is only a α percent chance of observing a standard deviate

$$Z_d = \frac{|d - \delta_0|}{\sqrt{V[d]}}$$

that is more than $z_{\alpha/2}$. Researchers occasionally calculate the probability of observing a normal deviate at least as large as Z_d; this is called the p-value. Alternatively, one may assess whether the hypothesized value lies inside the 95 percent confidence interval implied by the hypothesis.

Example. We now can reanalyze Lake's data as a difference of means (proportions). The effect of interest is the difference in success rates of Democratic regimes and Autocratic regimes. The estimated effect is $d = .81 - .43 = .38$. The variance of this estimate can be calculated by using the formula for independent samples and the formula for the variance of a Bernoulli: $v[d] = \frac{(.81)(.19)}{47} + \frac{(.43)(.57)}{74} = .006$.

A 95 percent confidence interval for the true effect is $.38 \pm 1.96\sqrt{.006} = .38 \pm .15$, or $(.53, .23)$. This is a very wide confidence interval.

Lake is interested in the hypothesis that there is no effect of regime type on success in war. If true, this hypothesis implies that both sorts of countries have the same proability of success, which would equal the overall rate of success in the sample ($= 70/121 = .579$). If this value is true, then the $V[d] = (.579)(1 - .579)(\frac{1}{47} + \frac{1}{74}) = .008$. The confidence interval implied by the null hypothesis is $.38 \pm 1.96\sqrt{.008} = .38 \pm .18$, or $(.56, .20)$. The hypothesized value lies outside this interval so we can reject the null hypothesis. To calculate the p-value we compute the probability of a normal deviate at least as large as $\frac{.38-0}{\sqrt{.008}} = 4.13$: $P(|Z| > 4.13) < .001$.

4.C. Experimental Logic and Design

We use the term effect to refer to a difference of conditional means. This is not a causal effect, however, and usually we are interested in measuring the extent to which one variable causes another. A simple definition of causality is that A causes B if in the presence of A, B occurs, and in the absence of A, B does not occur, all other things held constant (ceteris paribus). Causes may not be as absolute as this statement suggests. For example, A might cause a reduction in B, if B takes a continuum of values, and we would like to measure the extent to which A causes B. The important aspect about the definition is that all else is held constant.

Thinking statistically, we can state the definition in terms of conditional distributions. Suppose that there are individuals i and times t, and that a treatment variable X takes two values 1 and 0 and an outcome variable Y takes a continuum of values. The definition of causality is the difference in the conditional distributions when $X = 1$ and when $X = 0$. We say that the extent that X causes Y is captured by the difference in the random variables $(Y_{i,t}|X_{i,t} = 1) - (Y_{i,t}|X_{i,t} = 0)$. And the causal effect of X on Y is the exent to which the systematic component of the distribution differs when X changes:

$$\delta = E[(Y_{i,t}|X_{i,t} = 1) - (Y_{i,t}|X_{i,t} = 0)]$$

Immediately, there is a fundamental problem with the notion of causality. We cannot observe $X = 1$ and $X = 0$ for the same individual at the same moment of time. Consider the following problem. You would like to know whether a campaign commercial makes someone more likely to engage in some activity, such as to buy a product or to vote for a candidate. You cannot both show the commercial to someone and not show the commercial to someone. Another example, you might wish to know whether a medicine reduces the severity of headaches. You could do an experiment on yourself. But, you would have to take the medicine and not take the medicine when a given headache occurs in order to follow the definition of causality literally.

Commonly we observe the behavior of people who watch television commercials and of people who do not. Or, worse still the behavior of people who recall that they saw a commercial and of people who do not recall that they saw a commercial. We might also observe the behavior of people who take headache tablets and people who don't. Such studies are called *observational*.

It is difficult to infer causes from simple observational studies. Why? People who engage in a behavior X may be of a certain type of person, and we may simple measure differences in Y across types of people, rather than the effect of X on Y. For example, people who are very attentive to politics are more likely to recall that they saw a political commercial. More attentive people are also more likely to vote. The effect of recall of an ad on Y reflects attentiveness, not the effectiveness of commercials.

Such an effect is a *spurious* association. The ad does not really cause the behavior. The underlying attentiveness of the person causes the behavior.

Is there a way out of this conundrum?

> *Claim.* A controlled and randomized experiment is sufficient to guarantee an unbiased estimate of the causal effect.

What do we mean by control and randomization?

A *controlled experiment* is one in which the researcher determines which units (e.g., individual people at specific times) are assigned which values of X. For example, we could do an experiment in which we show some people a commercial and other people are shown no commercial. The group shown a commercial is called the *Treatment Group*, and the group not shown a commercial is shown a *Control Group*. The latter is the baseline behavior that would occur without the commercial and against which the effect of the treatment is measured. Importantly, the researcher determines who sees what – that is the essence of experimental control.

Experimental control is an excellent start, and it usually gets us "most of the way there." But not all of the way to an unbiased estimate. How are people to be assigned to the treatment and control? We could arbitrarily assign some people to watch a commercial and others to not watch a commercial. For example, we could recruit people to participate in our experiment at a shopping mall. The first person watches the commercial; the second does not; the third does; the fourth does not; etc.

Arbitrarily assigning people to groups might unwittingly introduce spurious associations. For example, suppose that couples come to the experiment and the man always arrives first followed by the woman. If we follow the approach above, then, one group will consist entirely of men and the other will consist entirely of women. The results of the experiment may, then, reflect differences in gender, which was introduced in the experiment by our assignment method, rather than the actual effect of the treatment.

Random Assignment, in addition to experimental control, guarantees an unbiased estimate of the true effect. Randomization involves using a device such as a coin toss or random number generator to assign individuals to experimental groups (i.e., treatment or control). In the case of the advertising experiment, we can introduce randomized assignment many ways. Here is one approach. The night before the experiment is to be done, the researcher takes the list of people scheduled to participate (say 100). The researcher then draws 100 random numbers from the first 100 numbers without replacement. The first number drawn is assigned to the first person; the second number to the second person; and so forth. Each person is assigned to the control group if the number drawn for them is odd and the treatment group if the number drawn for them is even. This approach randomly divides the list of 100 subjects into treatment and control groups.

How do we know *random assignment* and *experimental control* fix the spurious correlation problem? The sense in which randomized, controlled experiments work is that the Expected Outcome of the experiment equals the Theoretical Effect (δ). Let us generalize the definition of the experimental effect into the Average Causal Effect:

$$A.C.E. = \frac{1}{n} \sum_{i,t}^{n} E[(Y_{i,t}|X_{i,t} = 1) - (Y_{i,t}|X_{i,t} = 0)]$$

This treatment is due to Rubin (Journal of Ed. Stats. 1974).

I will give you a simple demonstration of this powerful idea.

For concreteness, consider the following problem. I wish to test whether route A or route B produces the faster commute home. There is one individual (me) and there are two time periods (today and tomorrow). I will commit to following one of two regimes. I can follow route A today and route B tomorrow, or I can follow route B today and route A tomorrow. This is the sample space of my experiment, and I control the assignment (though, regrettably, not the other drivers on the road). I could choose to follow A or B today on a whim. That would be arbitrary, and I might subconsciously follow a pattern that would bias my little experiment. For example, it looks cloudy, I'll take route A. But in the back of my mind I may have chosen route A because I know that there is more glare on route A on sunny days and thus a slower commute. This will obviously affect the outcome of the experiment.

Randomization involves tossing a coin to today. If the coin is heads, I follow Regime I – take A today and B tomorrow. If the coin is tails, I follow Regime II – take B today and A tomorrow. This slight change in my design is quite powerful. It means I will have an unbiased experiment. On each day I follow the chosen route and observe the dependent variable, the time of my commute.

The outcome of the experiment is defined as the difference in the dependent variable between group A and group B. That is the random variable is $Y(X_t = A) - Y(X_s = B)$, where Y is the time and X_j the route followed on day j, either t or s. Let $j = 1$ mean today and $j = 2$ mean tomorrow. Then the random variable has two realizations (or values):

$Y(X_1 = A) - Y(X_2 = B)$ if the coin toss is heads.

$Y(X_2 = A) - Y(X_1 = B)$ if the coin toss is tails.

The probability of observing the first realization is $p = .5$ and the probability of the second realization is $(1 - p) = .5$. The expected value of the random variable that is the outcome of the experiment is:

$$E[Y(X_t = A) - Y(X_s = B)] = \frac{1}{2}[Y(X_1 = A) - Y(X_2 = B)] + \frac{1}{2}[Y(X_2 = A) - Y(X_1 = B)]$$

$$= \frac{1}{2}[Y(X_1 = A) - Y(X_1 = A) + Y(X_2 = A) - Y(X_2 = B)]$$

The last expression is the Average Causal Effect. This shows that a randomized controlled experiment is unbiased: the expected outcome of the experiment equals the Average Causal Effect.

A second concern is not bias, but efficiency. This experiment is much too small and subject to idiosyncratic events that occur in traffic on days 1 and 2. It is really based on just one observation and is highly variable. A much larger sample is desired to get a more precise estimator. How large a sample size we need depends on how wide of an interval around the true effect we wish to estimate.

There is increasing use of experiments in political science and social sciences generally. And, even when we cannot perform a controlled experiment, the logic of experimentation provides a model for how we improve estimates through the careful design of studies. A good example is the literature on incumbency advantages.

Application: Incumbency Advantage. Let's consider 3 different study designs for estimating the incumbency advantage.

Design 1. Take the difference between the mean vote in all seats where there is an incumbent and the mean vote in all seats where there is no incumbent. In all 491 open seats from 1978 to 2000, the average Democratic vote margin was .518 with a standard error of .131. Among the 2013 cases where a Democratic incumbent ran, the mean Democratic share of the vote was .662, with a standard deviation of .108. Among the 1512 cases where a Republican incumbent ran, the mean Democratic share of the vote was .350, with a standard deviation of .088. The incumbency effect is .144 ± .016 among Democrats and .168 ± .016 among Republicans.

What might bias this estimate? Different sets of districts are involved, and places that are more likely to be open are more likely to be close, exaggerating the effect.

Design 2. Take the difference between the mean vote in seats that were open by previously controlled by a party and the mean vote in seats where an incumbent of that party runs for reelection. There are a large number of cases where districts change, so we lose these observations. In Democratic held seats, 1279 Democratic incumbents had average vote of .656 with a standard deviation of .105 and 134 open seat races had average Democratic vote of .550 with a standard deviation of .133. In Republican held seats, 1030 Republican incumbents had average Democratic vote share of .345 with a standard deviation of .09 and 128 open seat races had average Democratic vote of .437 with a standard deviation of .08. The incumbency effect was .106 ± .028 among the Democratic seats and .092 ± .019 among the Republican seats.

What might bias this estimate? Different people are running in different districts. If better candidates survive electoral challenges then Incumbents reflect a different pool of people than Open Seat candidates.

Design 3: Sophomore Surge. Take the difference in the vote between time 1 and time 2 for all incumbents who won in open seats in the previous election. Among the 155 cases where this is true for Democratic incumbents, the average increase in the Democratic incumbent's vote share was .04 with a standard deviation of .08. Among the 161 cases where this is true for Republican incumbents, the average increase in the Republican incumbent's vote share was .06 with a standard deviation of .09. Overall, the average increase in vote share was .05 with a standard deviation of .085. The confidence interval for the incumbency effect using this method is .05 ± .009.

5. Regression

Regression provides a general model for analyzing the conditional mean, $E[Y|X = x]$, of the joint distribution $f(y, x)$. Regression generalizes the concept of an effect to any type of variable X, not just binary. Regression also allows us to hold other factors constant in the analysis of the data (rather than the design), thereby lessening concerns about bias, and it provides a framework with which to use models to make predictions. As a result, regression is the foundation for most statistical research in the social sciences.

In the previous section, we defined an effect as the difference in the mean of Y across values of X, where X is binary. If X takes many values, the concept can be expanded to the change in the mean of Y given a unit change in X. That is, Y is a function of X and the effect of X on Y is $\delta = \frac{dE[Y|X=x]}{dx}$. Integrating with respect to X yields $E[Y|X = x] = \gamma + \delta x$. This is a linear representation of the effect of X on Y. We can generalize this further, making the effect itself variable, say $\delta(X)$.

Consider three examples. In an experiment relating advertising to behavior (consumption or voting), researchers wish to know how individuals respond to repeated advertisements. What is the expected effect of 1 ad, of 2, 3, 4, etc.? This relationship might be aggregated into a campaign production function to measure the returns to campaigning in terms of sales or votes. Suppose we performed an experiment in which we divided the participants into 5 groups and showed each group a number of ads – zero ads, one ad, two ads, three ads, and four ads. We then measured the attitudes and behaviors of the groups. A simple summary model of the effects of the ads on the participants is the linear model $E[Y|X = x] = \gamma + \delta x$, where x ranges from 0 to 4. The expected difference between someone who saw 4 ads and someone who saw no ads is $\delta 4$. One might further analyze each level separately to map out a response function. That is, one might measure the effect of an additional ad, given that the person has already seen 1, or seen 2, or seen 3, etc. Gerber and Green (APSR 2002) describe one such experiment. They find that the marginal returns to direct mail political advertisements are sharply decreasing. The difference in participation between those who receive 1 mailer and 2 mailers is larger than the difference in participation between those who receive 2 mailers and 3 mailers, and so on. After 6 mailers the marginal return is 0.

A second example is observational. What is the value representation? Many legislatures have unequal represenation (such as the US Senate), and before the court's eradicated such malapportionment in the mid-1960s, state legislatures commonly had inequalities in which some seats would have 20 times as many people (and thus 1/20th the representation) as other seats. Ansolabehere, Gerber, and Stewart (2002) use court-ordered redistricting to estimate how an increase in representation affects the share of funds an area receives from the state. They found that doubling a county's representation increased that county's share of state revenues 20 percent.

A third example derives the exact functional form from theory. Ansolabehere, Snyder, Strauss and Ting (2002) consider the division of cabinet portfolios under theories of efficient bargaining. From existing theories they derive the condition that any party asked to join a coalition government can expect a share of posts proportional to its share of "voting weights": $Y_j = cX_j$ if party j is a coalition partner. And, the party chosen to form a coalition government receives its proportionate share of the posts plus any surplus: $Y_f = (1 - \sum_j cX_j) + cX_f$ for the party that forms the coalition. Assuming that minimum winning coalitions form, they show that the expected division of cabinet posts is: $Y_i = F_i(1 - \frac{W+1}{2W} + cX_i) + (1 - F_i)cX_i$, where F_i is a binary variable indicating which party forms the coalition, W is the total voting weight of the parties, and c is the implied price of buying a partner with 1 vote. The function simplifies to a linear function: $Y_i = F_i(1 - \frac{W+1}{2W}) + cX_i$. This is a regression of each coalition members share of posts on their share of parliamentary "voting weight" plus a binary variable for the formateur. The regression estimates the price of a coalition partner, predicted by theories to range from 1 to 2, and the advantage of being formateur, predicted by various theoretical models to range from 1/2 to 0. Analyses of parliamentary coalition governments from 1945 to 2002 show that the coefficient c in this model is slightly larger than 1 and the coefficient on the variable indicating whether a party formed the government is .25, indicating an advantage to forming the government.

The second two examples involve observational analyses, which is the dominant form of study in social sciences. The experimental example is in many ways ideal, because we can be more confident that the experiment caused any effect, as opposed to some other factors.

The leap from experimental thinking to observational thinking is difficult. In fact, statistical methods for studying observational data emerged nearly a century after the same methods for experimental work. Importantly, the concepts and methods for studying observational data turned out to be the same as the experimental methods.

Physical experiments gave rise to the concept of regression. At the end of the 18th Cetury, Gauss and Laplace developed a simple elegant model to measure the underlying relationship between two variables as reflected in in experimental data. The idea was this. The relationship between X and Y is fixed but the parameters are not exactly known. The experimenter chooses the level of a variable, X, and then observed the outcome Y. Y is observed with some measurement error. In Gausses case astronomical observations were made and atomosphere introduced measurement error. Gauss and Laplace separately developed a method for extracting the underlying parameters from the observed data called least squares. Given fixed values of X, the problem was one of minimizing measurement error.

Sir Robert Galton, an English statistician in the late 19th Century, observed that the same statistical procedure applied to the measurement of two variables in social data, even when the data were not experimentally derived. Galton studied, among other phenomena, human genetics. He observed that the height of offspring was on average a bit lower than the mid-point of the height of the parents when the parents were above average. Also, the height of offspring was on average a bit higher than the mid-point of the height of the parents when the parents were below average. This was termed regression to the mean. He also observed that the relationship between parents' heights and offsprings' heights followed a jointly normal distribution.

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

The conditional mean of this distribution, it was further observed, is a linear function:

$$E[X_2|X_1 = x_1] = \alpha + \beta x_1,$$

where $\alpha = \mu_2 - \frac{\sigma_{1,2}}{\sigma_1^2}\mu_1$ and $\beta = \frac{\sigma_{1,2}}{\sigma_1^2}$. Using sample quantities to estimate these parameters yields exactly the same formula as Laplace and Gauss derived.[2]

How are the data most efficiently used to estimate the coefficients? How can we use estimated values to draw inferences or make predictions? Given the information a study can contain, we also seek to find ways to improve designs so as to get the most information out of the fewest cases (efficiency) and to guard against spurious relationships (bias).

5.A. Model

The regression model can be arrived at from many different approaches. As a start, consider the model of means presented in the last section: $Y_{i,j} = \mu_j + \epsilon_{i,j}$, where i indexes individual units and j indexes groups or values of X. Let μ_j be a linear function of values of X: $\mu_j = \alpha + \beta x_i$. Then, $Y_i = \alpha + \beta x_i + \epsilon_i$. Sometimes this is called a generalize linear model in which there is a "linking function" defining the mean plus an error term. This is way of thinking about regression tends to originate with experimentalists. We determine the values of X and perform an experiment in order to figure out how the mean of Y depends on values of X. The left over term is a random error component that is unrelated to X because of the structure of the experiment.

A more subtle idea is that we seek to estimate the conditional mean and conditional variance of the distribution of $Y|X$. Linearity is a good starting approximation for the function describing these means. Sometimes we might encounter data that are jointly normal, and thus, the function describing the conditional means is linear in x. Sometimes we can derive from a theoretical model a formula that relates empirically observable variables.
The basic linear regression model is a set of assumptions describing data:

[2] That is, substitute $\bar{x}_j$ for μ_j, s_j for σ_j, and $s_{1;2}$ for $\sigma_{1;2}$.

(1) A Linear Relationship defines how Y depends on X

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

Or, in the case of many different independent variables, X_1, X_2, etc.:

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$$

(2) $E[\epsilon_i] = 0$

(3) Independence of X and ϵ (no omitted variables): $E[X_i \epsilon_i] = 0$.

(4) No measurement error in X.

(5) "Spherical Distribution of Errors": (a) Constant error variance (homoskedasticity): $E[\epsilon_i^2] = \sigma_\epsilon^2$, and (b) No autocorrelation: $E[\epsilon_i \epsilon_j] = 0$ if $i \neq j$.

(6) ϵ follows a normal distribution.

Assumption (3) is critical. When this assumption fails, biases in estimates emerge, and are sometimes so severe that the results are non-sense. The most common problem is that the error in the regression model captures all factors not in the model, some of which may be systematic but unmeasured. Some of those unmeasured variables may be correlated with the included variables, X_1, X_2, etc. This assumption is extremely difficult to test and validate. It is, therefore, the assumption that requires the greatest thought and care in the *design* of any study.[3]

A carefully performed experiment estimates the true effect without bias. Randomization and control help ensure that assumption (3) is satisfied. Randomization and control (if there are no problems with compliance) mean that the level of X that a unit receives is not correlated with anything else.

Most of what we learn, however, we learn from observation. And, in observational studies, assumption (3) is less surely met. The reasonability of estimated parameters usually indicate

[3] Other possible problems emerge because of aggregation or simultaneous relationships between dependent and independent variables.

whether there is a severe violation of assumption (3). For example, in the field of criminology, it is well known that there is a positive correlation between crime rates and number of police on the streets. A regression of violent crime rates on number of police in major U.S. cities from 1975 to 1995 has an estimated regression line of $Crime = .08 + .13Police$. More police do not cause more crime.

There are likely two problems that violate Assumption (3). First, there are many omitted factors, such as the age distribution, income rates, drug use rate, and so forth. These must be held constant in order to estimate correctly the effect of increasing the number of police on the crime rate. Second, the relationship between crime and police may be simultaneous. A city experiencing a crime wave, whatever the cause, is likely to increase the number of police on the street. Many sociologists, political scientists, and economists have tried to tackle this problem. For a survey of research through the 1980s see Wilson Thinking About Crime. For recent innovative work, see Levitt (1999).

Assumptions (1) and (2) concern the functional form. I will develop the framework using linear models. A more complicated function may be required, though. For example, in Gerber and Green's advertising experiments, the functional form exhibits decreasing returns and is clearly not linear. Ideally, theoretical analyses, such as a game theoretic model or a conjecture from psychology, will guide us to the choice of functional form. The wrong functional form can be biased or inefficient, or both.

Fortunately, we can usually tell from the data whether the linear model makes sense, at least as an approximation. Four sorts of non-linear models are common – step functions (dummy variables), interactions among variables, transformation into the logarithmic scale, and quadratics. These are readily accomodated within the linear structure.

Dummy Variables. A step function, such as a difference of means or a shift in intercepts, can be modeled with the inclusion of a binary variable that equals 1 if the logical statement defining the shift is true and 0 otherwise. In regression analyses, variables that indicate such shifts or steps are called Dummy Variables. An example is the indicator of which party forms the government, F_i, in the coalition government analyses above.

Note: The difference of means consists of the regression of Y on X where X is binary.

Interactions. Interactions arise most commonly in psychological and behavioral models. As an example of an interaction, consider the psychology of advertising. An advertisement may be more effective among people who hold beliefs consistent with the message of the ad or who care most intensely about the issue or product in the ad. Among such people the effect of an additional ad might be quite strong (and possibly linear). Among all people not inclined to believe the ad or who don't care about the message, the effect is small.

Multiplicative Models and Logarithms. Logarithmic transformations are perhaps the most common. The logarithmic model makes a multiplicative model linear, and thus easier to analyze. Specifically, suppose that the true relationship among the variables of interest is as follows:

$$Y_i = AX_i^{\beta}u_i$$

Taking (natural) logarithms of both sides of this equation yields the following linear model:

$$log(Y_i) = log(A) + \beta log(X_i) + \log(u_i) = \alpha + \beta log(X_i) + \epsilon_i$$

This is a linear model, except that that scale of X and Y have been changed. The new scale is in terms of *percentages*. That is, each unit increase of $log(Y)$ represents a one percent increase in Y. We can see this from the derivative: $\frac{dlog(X)}{dX} = \frac{1}{X}$. If $X = 1$, a unit increase is a 100 percent increase in $log(X)$; if $X = 10$, a unit increase in X is a 10 percent increase in log(X); and so forth.

The slope coefficient in the multiplicative model transformed into the linear scale is interpreted as an elasticity. Specifcially, for a one percent increase in X there is a β percent increase in Y.

Polynomials. Other sorts of transformations are also common, especially quadratics. Polynomials, such as quadratics, are used to approximate a general function. For example, in the beginning of the course we used quadratics to approximate the returns to schooling and on the job experience. As we learned earlier, higher ordered polynomial terms can be included in the regression to capture an unknown curved function.

Assumption (4) holds that the measurement of X is made accurately. While X is a random variable, we must also take care not to introduce additional variation due to the instrument used to measure X. If we have a noisy measuring device, we will introduce random measurement error which will *tend* to bias estimates toward 0. In a bivariate regression that bias will surely arise; in a multivariate analysis the bias may have any sign.

In some problems, measurement error is a necessary evil. Proxy variables are often used in social sciences to measure a concept with an index or some variable that captures the concept. Proxies involve measurement error and thus produce bias in estimates. One interesting methodological question is when is a proxy variable worse than no variable at all? The concensus seems to be that it is always best to include a proxy when possible.

Assumptions (5) and (6) are less essential to the model. The do affect the efficiency with which we estimate the parameters of the data. Violations of these assumptions are readily fixed.

How we generalize the model and how we deal with violations of these assumptions are the challenges for the next course in this sequence. For the remainder of this course we will focus on the analysis of the model, assuming that the assumptions hold.

5.B. Estimation

5.B.1. Estimation Methods

There are three parameters to estimate in the simple (bivariate) regression model – the slope (β), the intercept (α), and the error variance σ_ϵ^2. We may stipulate many different estimation concepts. We can maximize the likelihood function. We can find the values that satisfy the method of moments. We could minimize mean squared error. All lead to the same answers, interestingly.

Estimation within the regression framework began with the idea of minimizing the error variance. This is the notion of Least Squares. It is the idea Laplace and Gauss developed. We often refer to regression as Ordinary Least Squares regression.

Define the sum of squared errors as follows:

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

This function is a bowl-shaped parabola in α and β.

To find the values of α and β that minimize this function, take the first derivatives with respect to each and set these equations equal to 0. The resulting equations are called the normal equations.

$$\frac{\partial S}{\partial \alpha} = \sum_{i=1}^{n} -2(y_i - (\hat{\alpha} + \hat{\beta} x_i)) = 0$$

$$\frac{\partial S}{\partial \beta} = \sum_{i=1}^{n} -2x_i(y_i - (\hat{\alpha} + \hat{\beta} x_i)) = 0$$

Solving the first equation for $\hat{\alpha}$ yields:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Substituting this result into the second normal equation yields:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Consider also the method of moments. There are two moments in the model: $E[\epsilon] = 0$ and $E[X\epsilon] = 0$, assumptions (2) and (3), respectively. Define the observed error (also called the residual) as $e_i = y_i - a - bx_i$. Assumption (2) implies $\sum_i e_i = 0$ and assumption (3) implies $\sum_i x_i e_i = 0$. These are the empirical moments that correspond to the theoretical momemts. Now let us find the values of a and b that satisfy these restrictions.

The first empirical moment is $\sum_i e_i = \sum_i (y_i - a - bx_i) = 0$. Solving for a yields $a = \bar{y} - b\bar{x}$.

The second empirical moment is $\sum_i e_i x_i = 0$. We can subtract $\bar{e}$ from e_i and $\bar{x}$ from x_i and the equation still holds. This yields $\sum_i e_i x_i = \sum_i (y_i - a - bx_i - (\bar{y} - a - b\bar{x}))(x_i - \bar{x}) = \sum_i (y_i - \bar{y} - b(x_i - \bar{x}))(x_i - \bar{x}) = 0$. Collecting terms and solving the equation for b:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The method of moments, then, yields the same results as least squares.

The final moment to estimate is σ_ϵ^2. It can be shown that $E[\sum_i^n e_i^2] = (n-k)\sigma^2$, where n is the number of observations and k is the number of parameters in the regression formula estimated (in the simple case k=2). Hence, $s_e^2 = \frac{1}{n-k}\sum_{i=1}^n e_i^2$.

One may further verify that maximum likelihood with normally distributed errors yields these estimators as well, with the caveat that the estimator of the error variance does not adjust for the degrees of freedom.

5.B.2. Properties of Estimates

The estimators a, b, and s_e^2 are functions of random variables because they depend on y. As such, they are themselves random variables. From sample to sample the values observed will vary. What are the distributions of the estimators of the regression parameters.

(1) The estimators are unbiased. $E[a] = \alpha$ and $E[b] = \beta$.

I will show this for b. We must consider the expected value of b conditional on the values of X. For convenience I will drop the conditional.

$$E[b] = E[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}] = E[\frac{\sum_{i=1}^n (x_i - \bar{x})(\alpha + \beta x_i + \epsilon_i - \alpha - \beta\bar{x} - \bar{\epsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2}]$$

$$= E[\frac{\sum_{i=1}^n (x_i - \bar{x})(\beta(x_i - \bar{x}) + (\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2}] = E[\beta + \frac{\sum_{i=1}^n (x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2}] = \beta,$$

assuming that ϵ and X are uncorrelated (assumption(3)). Note: This also means that if the two are correlated then the estimated regression slope may be biased.

(2) The variances of the estimators are:

$$V[b] = \frac{\sigma_\epsilon^2}{\sum_i (x_i - \bar{x})^2}$$

$$V[a] = \sigma_\epsilon^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}]$$

In deriving these results, we use the “homoskedasticity” assumption and the “no autocorrelation” assumption.

$$V[b] = E[(b-\beta)^2] = E[(\beta + \frac{\sum_{i=1}^n (x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2} - \beta)^2] = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 E(\epsilon_i - \bar{\epsilon})^2}{\sum_{i=1}^n (x_i - \bar{x})^4}$$

$$= \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$V[a] = E[(\bar{y} - b\bar{x} - \alpha)^2] = E[(\alpha + \beta\bar{x} + \bar{\epsilon} - b\bar{x} - \alpha)^2]$$

$$= E[(\bar{x}(\beta - b) + \bar{\epsilon})^2] = E[(\bar{x}^2(\beta - b)^2] + E[\bar{\epsilon})^2]$$

$$= \frac{\bar{x}^2\sigma_\epsilon^2}{\sum_i (x_i - \bar{x})^2} + \frac{\sigma_\epsilon^2}{n}$$

(3) The covariance of the estimated parameters is $Cov(a, b) = \frac{-\bar{x}\sigma_\epsilon^2}{\sum_i (x_i - \bar{x})^2}$.

(4) The parameter estimates follow a joint normal distribution (from the Central Limit Theorem).

Simulations help us further develop our intuitions about the distribution of the estimators. I simulated the distribution of the parameter estimates for a sample size of $n = 200$ and a regression $y_i = -.3 + .8x_i + \epsilon_i$. The values of *epsilon* were drawn from independent normal random variables with mean 0 and variance 1. Two different simulations were performed for the values of X. In each the distribution is assumed normal. In the first, the mean of X is 2 and standard deviation 1. In the second, the mean of X is 2 and the standard deviation is 3.

I used the following STATA code.

```
set mem 300m
set obs 200
forvalues i = 1(1)2500 f
    gen x_‘i’ = invnorm(uniform())
    gen y_‘i’ = -.3 + .8*x_‘i’ + invnorm(uniform())
    quietly regress y_‘i’ x_‘i’
    matrix list e(b)
    g
```

This returned 2500 values from the distribution of b and a for a regression with sample size of $n = 200$.

Simulated values of a had an average of -.302 and a standard deviation of .160. Simulated values of b had an average of .801 and a standard deviation of .073. From the formulas, we expect means values of a and b of -.3 and .8, respectively, and standard deviations (standard errors) of .158 and .071.

A further feature of the estimator that we have derived deserves emphasis. The least squares/method of moments estimator has the smallest variance of all possible linear and unbiased estimators. That is, any other way of adding the data together to form estimates of a, b, and σ_ϵ^2 that is unbiased will have higher variance. Least squares uses the data most efficiently. This result is general and is called the Gauss-Markov Theorem

Consider the following alternative estimator. Choose the smallest value of X, x_s and the largest value of X, x_L. Observe the corresponding values of Y, y_s and Y_L. We can estimate the slope as

$$\tilde{b} = \frac{y_L - y_s}{x_L - x_s}$$

This esimator is unbiased. Of note, it is not unbiased if we choose the highest value of Y and the lowest value of Y and observe the corresponding X's. To show unbiasedness evaluate the expected value of $b|X$.

$$E[b|X] = E\left[\frac{y_L - y_s}{x_L - x_s}|X = x\right] = E\left[\frac{\alpha + \beta x_L + \epsilon_L - \alpha - \beta x_s - \epsilon_s}{x_L - x_s}|X = x\right]$$

$$= E\left[\frac{\beta(x_L - x_s) + (\epsilon_L - \epsilon_s)}{x_L - x_s}|X = x\right] = \beta + E\left[\frac{\epsilon_L - \epsilon_s}{x_L - x_s}|X = x\right]$$

The last term in the equation is 0 because $E[\epsilon|X] = 0$. (This is the assumtion violated by conditioning on Y.)

Now we may consider the variance of $\tilde{b}$

$$E[(\tilde{b} - \beta)^2] = E[(\beta + \frac{\epsilon_L - \epsilon_s}{x_L - x_s} - \beta)^2|X = x] = \frac{2\sigma_\epsilon^2}{(x_L - x_s)^2}$$

which is larger than $\frac{\sigma_\epsilon^2}{\sum_i (x_i - \bar{x})^2}$.

5.C. Inference

Inference about regression models follows the framework developed for differences of means. Subtleties arise when we consider more than one estimated coefficient, or when we compare regression models. We begin with confidence intervals and tests concerning a single coefficient.

1. Inference about Coefficients

In developing the basic tools of inferences, we will consider two examples: the relationship between seats and votes in a two-party system and the relationship between a party's share of cabinet poses and its share of voting weights in coalition governments. We have discussed the latter example already.

A first sort of inference is a confidence interval for a coefficient. The estimates of a and b and their distributions allows us to construct confidence intervals readily. Since the coefficient estimates are sums of random variables (sums of ϵ's), we know that the distribution of a and b tends to the normal distribution, regardless of the underlying distribution of the data.

Consider again the estimated relationship between a party's share of cabinet poses and its share of voting weights in coalition governments. Holding constand whether a party is formateur or not, the estimated effect of voting weights is 1.16 with a standard error of .07. A 95 percent confidence interval for this coefficient is $1.16 \pm (1.96)(.07) = [1.30, 1.02]$. Similarly, a 95 percent confidence interval the estimated coefficient on formateur is $.145 \pm (1.96)(.03)$.

2. Inference about Models

Theoretical models and arguments have concrete predictions. If the argument is right, regression coefficients ought to equal specific values.

Competing models of coalition formation carry specific predictions. Gamson's law holds that a is not distinguishable from 0 and b, the coefficient on voting weight, is not distinguishable from 2; the Baron Ferejohn model holds that $a = 0$, $b = 1$ and c, the coefficient on the dummy variable for Formateur, is not distinguishable from .5.

Another example is the "Cube Law." At the beginning of the 20th Century, statisticians observing English elections posited a "law of cubic proportions" describes the rate at which votes are translated into seats. It is this:

$$\frac{S}{1-S} = \left(\frac{V}{1-V}\right)^3$$

James March developed a regression model wihtin which to nest this model. Taking logarithm of both sides of the cube law yields a linear relationship: $log(S/1 - S) = 3log(V/1 - V)$. Define $Y = log(S/1 - S)$ and $X = log(V/1 - V)$. The cube law predicts that the estimated coefficients of the regression $Y = \alpha + \beta X$ should not be distinguishable from $a = 0$ and $b = 3$. The coefficient a has since been renamed the "bias." If $a \neq 0$, then when two parties split the vote evenly, one party receives significantly more than half the seats.

All hypothesis tests take the form of the probability that the coefficients of interest deviate from an hypothesized value.

For a single coefficient, we construct the standarized value of the coefficient if the hypothesis is true:

$$t = \frac{b - b_0}{SE(b)},$$

where b_0 is the hypothesized value and $SE(b)$ is the standard error of b. If the number of cases is relatively small (say less than 100), then the statistic follows the t-distribution with $n - k$ degrees of freedom, wher k is the number of variables in the model. If the number of cases is relatively large, the normal distribution approximates the t-distribution well and we can calculate the probability of observing a deviation at least as large as t with reference to the standard normal.

For example, we may ask whether c differs from 0 in the coalition government analyses, that is whether there is a significant formateur advantage. The test statistic is $\frac{(.145-0)}{xxx} = .xxx$. The probability of observing a standardized deviation of c from 0 that is at least this large is smaller than .01, so we reject the null hypothesis at the 99 percent level. Note this is the same as asking whether the hypothesized value lies inside the confidence interval.

When hypotheses involve multiple coefficients, one cannot test each coefficient in isolation. The estimated coefficients $a, b, c, ...$ are jointly normally distributed, but they are not independent. Hence, the probability of observing a given deviation in one coefficient and a deviation in another coefficient does not generally equal the product of the coefficients. We could construct a joint confidence interval, which would consist of the ellipse defined by $f(a, b) = .05$. Hypothesized values of a and b inside this ellipse are supported by the data, but values outside the data are not supported.

An alternative approach is to consider the loss of fit that results by maintaining the hypothesis. When the hypothesis is maintained or imposed on the data, the amount of variation in y not explained by X is $\sum_i u_i^2$. When the hypothesis is not maintained, the amount of variation in y not explained by X is $\sum_i e_i^2$. Let J be the number of parameters constrained by the hypothesis and k the total number of parameters. The percentage change in fit from the imposition of the hypothesis is:

$$F = \frac{(\sum_i u_i^2 - \sum_i e_i^2)/J}{\sum_i e_i^2/(n-k)}$$

If there is a significant loss of fit from imposing the hypothesis then the amount of unexplained error will be large.

This formula, it turns out, is identical to calculating the sum of squared deviations of each parameter estimate from the value implied by the hypothesis divided by the variance of that sum of squared errors. The square root of this formula is a general form of the t-distribution.

To determine whether the observed loss of fit could have arisen by chance, we calculate the probability of observing a value of the random variable defined by the F-statistic that is at least as large as the observed value of the statistic. If that probability is very small then the observed loss of fit is unlikely to have occured by chance.

The F-statistic follows an F-distribution. An F-distribution, as mentioned earlier in the course, is the distribution that arises from the ratio of squared normals, that is the ratio of χ^2 distributed random variables. Because the sum of χ^2 random variables is also χ^2 we can construct many F-distributions depending on the number of χ^2 in the numerator and the number in the denominator. In the case of the statisic above, there are sum of J independent squared normals in the numerator and $n - k$ independent squared normals in the denominator. (Why only J independent normals in the denominator? Because we use $n - k - J$ when we maintain the hypothesis and $n - k$ when we don't. Taking the difference between the sum of squared errors leaves J pieces of information free.)

In STATA, we can implement this test by using the **test** command following a regression as follows

```
reg y x1 x2
test x1=k1
test x2=k2, accum
test _cons=a0, accum
```

The test command performs an F-test for each variable, one at a time. We consider multiple restrictions on coefficients using **accum**.

The Cube law implies both that $b = 3$ and $a = 0$. To test this hypothesis for England (where it originates), we estimate the regression model proposed by March. Data consist of the Conservative party's share of the seats and share of the votes in elections from 1927 to 2002. The estimated intercept and slope are -.04 and 2.65, with standard errors .06 and .23, respectively. Imposing the cube law on the data implies that $u_i = log(\frac{S_i}{1-S_i}) - 3log(\frac{V_i}{1-V_i})$. The sum of these residuals squared and divided by 2 is the numerator of the F-statistics. Without imposing the values of a and b, we estimate the regression model and use the mean-squared error of the residuals for the denominator. The F-statistic is 1.81 for this problem, which follows and F-distributoin with 2 and 15 degrees of freedom. The probability of observing such a deviation were the hypothesis true is .20, so the cube law is supported in the data.

Note: The test of a single parameter is somewhat misleading. The test of whether the slope equals 3 is $t = \frac{2.65-3.00}{.234} = 2.82$, which is unlikely to have arisen by chance. But that is a partial test.

Consider Gamson's law and the Baron-Ferejohn bargaining model. Gamson's law implies that in a regression of shares of posts on shares of weight plus a formateur the coefficient on the dummy variable for the formateur should equal 0 and the coefficient on voting weights should equal 2. The Baron-Ferejohn model implies that the coefficient on the dummy variable for the formateur should equal .25, the coefficient on the share of voting weight should equal 1, and the intercept should equal 0.

We can test each of these models separately. The F-test for the appropriateness of Gamson's law is 273.4, with 2 and 244 degrees of freedom. The probability of observing a deviation at least this large is smaller than .001. This means that the observed deviations from expectations are quite unlikely to have occured by chance if that theoretical model captured bargaining well. The F-statistic testing the Baron-Ferejohn model is 8.73, which is also unlikely to have arise by chance if the model is exactly right.

Neither model fits the data sufficiently well that we would readily accept it as the right model of parliamentary bargaining. The signficant formateur advantage deviates from Gamson's law, but it is not large enough to indicate that the Baron-Ferejohn model is right. What can we conclude from tests showing both models are wrong?

3. Inference about Predictions

A final sort of inference of importance concerns predicted values. Policy analysis commonly uses data analyses to generate predictions. For example, if the economy grows at a certain rate, then tax revenues will grow by that amount and government revenues will either fall short of or exceed the amount budgeted. Political scientists often make forecasts about elections based on surveys or models in which the aggregate vote is predicted by economic growth and international conflict. One rule of thumb from such models is that if growth is 3 percent or more, the president will be reelected.

Let x_0 be a specific value of interest (such as 3 percent growth). Suppose we have estimated a regression model $\hat{y} = a + bx$. The *point prediction* is $\hat{y_0} = a + bx_0$. The predicted value is $y_0 = a + bx_0 + e_0$, the point prediction plus the residual.

How much uncertainty is associated with predictions? The uncertainty about predictions is of two sorts. First, we are uncertain because of the random variation inherent in a single event, such as a specific election. Second, we are uncertain about the regression line. This uncertainty is the variance of the predicted value:

$$\begin{aligned} V(y_0) = V(a + bx_0 + e_0) &= V(a) + V(b) - 2Cov(a, b) + V(e_0) \\ &= \sigma_\epsilon^2 (\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}) + \frac{\sigma_\epsilon^2}{\sum_i (x_i - \bar{x})^2} - 2\frac{\sigma_\epsilon^2 \bar{x} x_0}{\sum_i (x_i - \bar{x})^2} + \sigma_\epsilon^2 \\ &= \sigma_\epsilon^2 [1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}] \end{aligned}$$

A couple of aspects of the prediction error are noteworthy. First, it never shrinks to zero as the sample size becomes large. The lower bound is σ_ϵ^2. This is because prediction ultimately concerns a single event. Second, the farther the value of x_0 from the mean of $\bar{x}$, the worse the prediction. An "out of sample" prediction is based on a value of X farther away from the mean than the smallest or largest value of X in a sample. Such predictions are extremely uncertain.

We may use the $V(y_0)$ to form a confidence interval for predictions in the same way as we form confidence intervals for means: $\hat{y_0} \pm 1.96\sqrt{V(y_0)}$.

For example, a simple model of the last 13 presidential elections uses income growth to predict presidential election votes. The regression has an intercept of 48.4 and a coefficient on income of 2.41. The average growth rate is 1.6, the estimated error variance is 19.26, and the sum of squared deviations of X is 72.62. If income growth is 3 percent this year, Bush is predicted to win 55.3 percent of the vote. But the confidence interval on this prediction is 9.06.

A more sophisticated model includes an indicator for incumbency and an indicator for war, as well as income. The estimated regression is *Vote* = 47.3 + 2.08 *Income* +5.74

Incumbent -6.04 *War*. Assuming continued military engagements in Irag and Afghanistan, growth of 1.6 percent implies $\hat{Y}_{2004} = 50.8$, growth of 2 percent implies $\hat{Y}_{2004} = 51.6$, and growth of 3 percent implies $\hat{Y}_{2004} = 53.7$.

It is easy to have a false sense of confidence about these predictions because they are hard numbers. Average growth and this is a very titght race. Adding war and incumbency improves the model, but the precision is poor for a predictive model. The MSE is 4, so the standard error of the predicted values are approximately 4.2.

5.D. Design

I would like to highlight four lessons to close the semester.

First, think about social phenomena as random variables. It is rare that we have truly deterministic theoretical models, and the state of knowledge in political science and other social sciences is such that there is a great deal of variation that we do not understand. As a conceit we treat such variation as random. It is our hope to capture important (large) systematic variation.

Second, the outcomes of studies are themselves random variables, depending on what cases were studied and how the researcher measured the variables of interest. It is through the accumulation of knowledge across many studies that we learn.

Third, think multi-dimensionally. It is easy to seize on a single cause for phenomena. That reflects a basic commitment to parsimony. However, most social phenomena are predicted by many factors and, it is thought, have multiple causes. Observational studies that do not capture the important causal factors are bound to be biased. The evolution of understanding and knowledge occurs when ideas are subjected to analyses that introduce successive improvements in design to capture these .

Fourth, think backward. Good research design begins with conjectures – possible findings and the conclusions one might draw from them. Statistics provides you with a very useful framework for thinking through design problems. We must guard against false negatives and false positives in analyzing data relevant to a given conjecture. Doing so involves sampling from the population in ways that give you the greatest efficiency. That usually involves a somewhat large number of observations, but not always. Choosing very different values of the independent variable in an experiment or observational study yields the highest precision (and allows for smaller samples).