INTRODUCTION TO STATISTICS FOR POLITICAL SCIENCE: 2. Math Review

Stephen Ansolabehere
Department of Political Science
Massachusetts Institute of Technology

Fall, 2003

2. Mathematics Review

Models are mathematical representations of a behavior or phenomenon of interests. There are many different ideas about how to go about building a model. In political science today, game theory is the foundation for much formal modeling (e.g., bargaining models of the division of resources in legislatures). There are other traditions, such as dynamic models in which equilibria are "steady states" (e.g., Richardson's war model).

To be able to read and understand modeling requires a degree of familiarity with basic mathematical concepts. We will review them here. My review is in no way meant as a substitute for more in depth and rigorous mathematical instruction, which I strongly recommend.

We will begin this review with some basic algebraic ideas and then move onto concepts from Calculus.

1. Variables.

In algebra we use capital letters to denote a phenomenon or concept which can be represented using a metric, such as continuous numbers or counting numbers, and has a defined range. That range is specified by bounds, such as 0 to 100 or from $-\infty$ to $+\infty$. We will represent variables with capital letters usually from the end of the alphabet, e.g., X, Y, or Z. These are just names of the variables. We will represent a specific value (in the abstract) as a lower case letter, e.g., x, y, or z.

Several different sorts of variables arise commonly. One important distinction is between *continuous* and *discrete* variables. A continuous variable takes any value along a continuum of numbers. A discrete variable takes values that are separated by discrete intervals, such as different groups of people or objects (including gender, race, etc.) or ordered numbers (including ranks and counts).

An example of a variable is Income. The concept of Income is widely discussed in social sciences – for example, national income or gross domestic product, and income inequality. When we set out to measure income we must be more specific about units and the type of income. Often we deal with wage income, but there is also income from investments, inheritance, and other sources to be accounted for. There is household income and individual income. Also, income in one year has a different value or purchasing power than income in another year. Often researchers deflate variables using an index of prices to make data comparable over time. Usually income is treated as a continuous variable because the intervals of a penny or less are too small to consider meaningful.

2. Sets.

A set is a collection of things or numbers. Sets are widely used in statistics. We will use capital letters at the beginning of the alphabet to represent sets, e.g., A, B, and C. An element is the simplest unit in a set – a unique individual in a set that cannot be further reduced. We denote elements with lower case letters, sometimes with subscripts, e.g., a_1, a_2, ... a_n are elements of A.

A set is often defined as a set of numbers. We use braces to mark the set and a bar — to separate the typical element from the set of numbers in the set. For example the set of all real numbers between 0 and 100 is written $A = \{x|X > 0 and X < 100\}$.

Two special sets are the empty set, which has no elements and is denoted ϕ, and the universal set, which contains all possible elements and is denoted - .

Three important operations on sets are unions ($\cup$), intersections ($\cap$), and complements. The union of two sets, written $A \cup B$, is the set of elements in both sets. The intersection of two sets, written $A \cap B$, is the set of points in both sets. The complement of a set, written $A^c or \overline{A}$, is the set of points not in the set.

For example, let $A = \{x|X > 0 and X < 100\}$ and $B = \{x|X > 50 and X < 500\}$. Then, $A \cup B = \{x|X > 0 and X < 500\}$ and $A \cap B = \{x|X > 50 and X < 100\}$. $A^c = \{x|X \leq 0, X \geq 100\}$.

If A and B contain no elements in common, then $A \cap B =$.

A final set concept is conditioning. Sometimes we might want to restrict the universe to a smaller set. We write this as (B — A). Conditioning amounts to partitioning the universe into the set in A and the set not in A and then Suppose that we wanted only to consider the set of cases where a certain condition holds, for example $A = \{x|X > 0\}$. We may then want to consider a new set only under the condition that A holds. For example, suppose that $B = \{x|X > -50 and X < 50\}$. Conditioning is very closely related to the intersection

The inference problem is a clear example of conditioning. One of two states of the world hold. The hypothesis is true or the hypothesis is not true. If the hypothesis is true, then if we decide it is false we commit an error. If the hypothesis is fale, the we commit another sort of error when we decide that the hypothesis is true. These two types of errors are called "false negatives" and "false positives."

3. Functions.

Functions map one set of numbers into another set of numbers. One may think of the function as having one variable as an input and another variable as an output. We will use a lower case letter to denote a function and parentheses to denote the argument, e.g., y = f(X). Functions often include constatnts (numbers) as well as variables. For example, the formula for a straight line is $Y = a + bX$. The constants a and b are called *parameters*.

Usually we reserve the term function for a specific sort of mapping – one for which an X returns one y. It is possible for many x values to return the same y, but not for many y from any one x. A circle, then, is not a function.

Generally, we will use Y for the output variable and X for the input variable. The output variable is called the dependent variable and the input the Independent variable.

The inverse of a function reverse the mapping of X to Y and becomes Y to X. We write this as $X = f^{-1}(Y)$.

A. Counting Functions

Two simple functions useful in analysis using sets are indicator functions and counting functions. An indicator function returns the value of 1 if a number is in a set and 0 otherwise.

For example, an indicator function may equal 1 if a person is female and 0 if a person is male. If we let j index elements in a set, then we may write the indicator of whether an individual is in A as $I_j(A = 1)$.

A counting function returns the number of elements in a set: $n(A)$. This equals the sum of the indicator functions of A for all individuals in the set. That is $n(A) = \sum_{j=1}^{n} I_j(A = 1)$. For a finite set the total number of elements $n = n(- \ )$. And, $n(\phi) = 0$. Also $n(A \cup B) = n(A) + n(B) - n(A \cap B)$.

The basic rule of counting involves enumerating the number of ways that a sequence of events can happen or tasks can be done. Suppose that a series of tasks are performed. The first task involves N_1 possible actions; the second involves N_2 possible actions; and so on. A sequence k actions can be performed $N_1 N_2 N_3 ... N_k$ different ways.

Two important counting functions emerge from this rule.

1. Factorials

If n is a positive integer, then the product of the integers from 1 to n is called n-factorial and is written $n!$. Note: $0! = 1$. $n!$ is the number of ways that n numbers can be ordered, or the number of ways that n tasks can be performed when there are n possible actions to take without repeating one of the actions. Suppose that we want to arrange n objects in an order. I choose one (arbitrarily) to begin a sequence. There are n possible choices; I choose one and set it aside. Then, I choose another object from the remaining (n-1); there are n-1 possible choices. If I continue in this vein I could have made the arrangment or order $n! = n(n-1)(n-2)(n-3)...21$ possible ways, or taken n! possible actions.

2. Binomial Coefficients

If we perform the same sort of task but do not care about the order, we will arrive at a different counting function. If we form a subset of size k from the set of n objects, without regard to the order, then the number of ways that could happen equals the number of ways the n objects could be arranged divided by the number of ways they could be divided into a group of size k and a group of size $n-k$. That is:

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

An example of the binomial coefficients is Pascal's Triangle. Consider the coefficients on each of the terms in exapansions of $(x+1)^n$. When $n=1$, $(x+1)^0 = 1$, so the coefficient is 1. When $n=1$, $(x+1)^1 = x+1$, so the coefficient on X is 1 and on 1 is 1. When $n=2$, $(x+1)^2 = x^2+2x+1$, so the coefficients are 1, 2, 1. When $n=3$, $(x+1)^3 = x^3+3x^2+3x+1$, so the coefficients are 1, 3, 3, 1. Each coefficent can be calculated using the binomial coefficient, where n is the power of the entire product and k is the exponent of any single term, e.g., $k=2$ for x^2.

The distribution of binomial coeffients is the core mathematical idea behind the Central Limit Theorem. For large enough n, the coefficients follow a bell-shaped curve.

B. Continuous Functions

We will use several types of continuous functions: linear, polynomial, logarithmic, and exponential.

Linear:

$$Y = a + bX$$

This is the formula for a straight line. The inverse of this function is:

$$X = \frac{Y-a}{b}.$$

Also, we may have several X variables, which gives the formula for a plane:

$$Y = a + bX_1 + cX_2.$$

Polynomials:

$$Y = a_0 + a_1 X + a_2 X^2 + a_3 X^3 + ...$$

The most common polynomial we will use is the quadratic.

An example is the spatial representation of ideological preferences. Let X be a party's announced policy or ideology and let θ be a voter's most preferred policy some point along the X-axis. Then, the voter's utility from party X winning is

$$U(X, \theta) = -(X - \theta)^2$$

This representation captures a bit of the spirit of modeling. It is a mathematical representation of some key ideas. It does involve some restrictions, such as symmetry.

Inversion of polynomials is often complicated. A simple polynomial of the form $Y = X^b$ is readily inverted using the bth-root. To isolate X, raise both sides of the equation to $1/b$: $X = (Y/a)^{1/b}$. A function with 1/2 in the exponent is the square root, 1/3 the cube root, and so forth.

Finally, negative exponents mean that an operation is in the denominator of a ratio. $Y = X^{-b} = 1/X^b$.

Example: The Cube Law. In the study of electoral systems, scholars and constitutional designers distinguish two sorts of systems. Proportional representation systems set the share of seats a party receives in the parliament equal to that party's share of the national vote won during the election. A party that wins 20 percent of the vote receives 20 percent of the seats. Plurality rule or first-past-the-post systems set up districts within the country, and parties or candidates win seats if they win a plurality of the vote in a district. A party's share of the national vote does not translate immediately into a share of the seats. However, patterns that have arisen in elections over time have given rise to a particular relationship, the law of cubic proportions (see Edgeworth 1898 and Kendall and Stuart 1950).

Let V be the proportion of votes won nationally and S be the proportion of Seats received. Then,

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

[End of Lecture 2]

Exponential:

$$Y = b^X,$$

where b is a number, such as 2 or 10.

Growth problems take an exponential form. For example, if each generation produces 2 offspring per individual, then after n generations there will be 2^n individuals. With such a birth process, after 10 generations there will be 1024 individuals in the population. An important class of models in the study of populations are birth-death models. In the study of a country's population one must also add in migration.

Exponentials have some special properties. $b^0 = 1$. Also, $b^X + b^Z = b^{X+Z}$. (A special base is $e = 2.71828....$, which we will derive below.)

Logarithm:

$$X = log_b(Y)$$

This function is defined as the value of X such that the number Y equals b to the X. If b is 2 and Y is 4, then X = 2. The $log(0)$ is not defined. It approaches $-\infty$.

Logarithms have some special properties. $log(1) = 0$. Also, $log(ZX) = log(Z) + log(X)$, and $log(Z/X) = log(Z) - log(X)$. From this second property it follows that $log(X^b) = log(X) + ... + log(X) = blog(X)$.

Logarithms are very widely used in data analysis because they convert a multiplicative relationship into a linear one and linear models are easier to work with. Assume $Y = AX_1^b X_2^c$. Then, $log(Y) = log(A) + blog(X_1) + clog(X_2)$. Note: if one of the variables takes the value 0, then the function is undefined because $log(0) = -\infty$.

Example: Cobb-Douglas production function. Suppose that there are two inputs used to produce a good, Y. The inputs are capital, K, and labor, L. The amount of Y produced with any K-L pair is $Y = AK^\beta L^\alpha$, $\alpha > 0$ and $\beta > 0$.

We will mix and match and combine these functions. Of particular note, the function for the normal curve combines the exponential and the quadratic. If Z is a standard normal, then Z ranges from $-\infty$ to $+\infty$ and

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

4. Limits.

Limits link continuous and discrete problems. This is the fundamental concept on which calculus and analysis is based and it is a fundamental tool in statistics, as we define probabilities as long-range frequencies. A few examples help build your intuitions and facility with this idea.

Consider a sequence of numbers $a_n = 1/n$. This sequence is $1, 1/2, 1/3, 1/4, 1/5, ...$ This sequence has a lower boundary – it is always larger than 0, but the set does not contain the point 0. However, as n becomes arbitrarily large the sequence becomes arbitrarily close to 0. We can choose any n we like, and as we choose larger values of n we approach 0 from above. We say that the limit of this sequence is 0.

Consider a second sequence of numbers $a_n = \frac{n}{n+1}$. This sequence is $1/2, 2/3, 3/4, 4/5, 5/6....$ This sequence also has a boundary, 1. The set is always lower than 1 and it becomes arbitrarily close to 1 for arbitrarily large values. We may actually derive the result for this case from the first case. Rewrite the sequence as $a_n = 1 - \frac{1}{n+1}$. The first part of the expression is always 1 for any n. The second part of the expression, we learned in the first example, becomes arbitrarily close to zero. Hence, the sequence in this example approaches 1 as n becomes very large.

These two examples exhibit both the idea of the limit and one method for evaluating limits. The method is to divide the sequence into a part that is a constant and the remainder and show that the remainder vanishes. Another method for evaluating limits is to establish boundaries that squeeze a sequence to a specific number.

A. Definition

A more precise definition of a limit is as follows. Let $\{a_1, a_2, a_3...\}$ be a sequence of numbers and let l be a number. We say that l is the limit of this sequence if for any small positve number ϵ, there is a positive integer M such that for all $n \geq N$, a_n is in the interval around l that is smaller than ϵ:

$$|a_n - l| < \epsilon$$

We write this as $lim_{m\to\infty} x_m = b$ or $x_m \to b$.

Limits are easily manipulated algebraically. Here are some useful properties of limits:

1. Limits of sums are the sums of limits: $\lim_{n\to N} a + bn = a + bN$.

2. Limits of powers are the powers of limits: $\lim_{n\to N} n^b = N^b$.

3. For two functions, the limit of the sum (or difference) is the sum (or difference) of the limit. Consider two functions $f(n)$ and $g(n)$, $\lim_{n\to N} f(n) + g(n) = f(N) + g(N)$.

4. The limit of the product (or ratio) of two functions is the product (or ratio) of the limit: $\lim_{n\to N} f(n)g(n) = f(N)g(N)$.

5. It follows from 1, 2, and 3 that the limit of a polynomial is the polynomial of the limit: $\lim_{n\to N} a_0 + a_1 n + a_2 n^2 + a_3 n^3 + ... = a_0 + a_1 N + a_2 N^2 + a_3 N^3 + ...$

B. Three Important Limits.

1. The number $e = 2.71828....$

$$e = \lim_{n\to\infty} 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + ... + \frac{1}{n!} = \lim_{n\to\infty} \sum_{n=0}^{\infty} \frac{1}{n!}$$

Another limit that equals e is

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n$$

Approximating e	
n	Approx. e
1	2
2	2.25
4	2.4414
10	2.59374
100	2.704814
1,000	2.7169239
10,000	2.17181459
100,000	2.17826824
10,000,000	2.178281693

2. The number $\pi = 3.14159....$ Wallis' product (see Courant, vol. 1, pages 223 - 226) is the ratio of the product of squares of n -1 even numbers and n-1 odd numbers. Specifically:

$$\frac{\pi}{2} = \lim_{n\to\infty} \frac{2}{1}\frac{2}{3}\frac{4}{3}\frac{4}{5}\frac{6}{5}...\frac{2n}{2n-1}\frac{2n}{2n+1} = \lim_{n\to\infty} \frac{2^2 4^2 6^2 ...(2n-2)^2}{3^2 5^2 7^2 ...(2n-1)^2} 2n$$

Algebraic manipulation reveals allows us to expresses π as a function of the factorial $n!$. Taking squareroots of both sides and multiplying top and bottom by $246...(2m-2)$, we get

$$\sqrt{\frac{\pi}{2}} = \lim_{n\to\infty} \frac{2 \times 4 \times 6... \times (2n-2)}{3 \times 5... \times (2n-1)} \sqrt{2n} = \lim_{n\to\infty} \frac{2^2 4^2 6^2 ...(2n-2)^2}{(2n)!} \frac{\sqrt{2n}}{2n}$$

This can be rewritten as

$$\sqrt{\pi} = \lim_{n\to\infty} \frac{2^{2n}(n!)^2}{(2n)!\sqrt{n}}$$

The term inside the limits gives a handy formula for approximating π.

Approximating $\sqrt{\pi} = 1.77245$	
n	Approx. $\sqrt{\pi}$
1	1
2	1.8856
4	1.8475
10	1.79474
100	1.774671
1,000	1.772675
10,000	1.772476

3. Stirling's formula. Many statistical problems involve $n!$, which is hard to calculate or work with in analytical formulae. Stirling's formula provides an approximation that is used in proving basic statistical results, such as the Central Limit Theorem for Binomials. Stirling's Formula is this:

$$n! \rightarrow \sqrt{2\pi} n^{n+.5} e^{-n}.$$

or $log(n!) \approx nlog(n) - n + .5log(n) + log(\sqrt{2\pi})$.

The formula can be derived by considering how well $log(n!)$ approximates the continuous function $y = log(x)$. The area under the curve $y = log(x)$ from 1 to n is $nlog(n) - n + 1$. One may approximate the area under $y = log(x)$ using the sum of trapezoids with base from n to $n+1$ and height defined as $log(n)$ and $log(n+1)$. The sum of the areas of these trapezoids is $log(2) + log(3) + log(4)... + log(n) - \frac{1}{2}log(n) = log(n!) - \frac{1}{2}log(n)$. Hence, $nlog(n) - n + 1 \approx log(n!) - \frac{1}{2}log(n)$. Taking exponents of both sides, yields $n^n e^{-n} e \approx n!/\sqrt{n}$, or $n! \approx n^{n+.5} e^{-n} e$. Further analysis reveals that the approximation can be improved further using $\sqrt{2\pi}$ instead of the last e (because the continuous function lies above the trapezoids constructed to approximate the area under the curve). Note, however, the similarity between $e = 2.71828$ and $\sqrt{2\pi} = 2.506628$.

C. Compounding and Interest Rates

You invest A dollars at an interest rate r. If the interest payment is made at the end of one year then you receive $A(1+r)$. What if the bank pays you interest after 6 months and then you can automatically reinvest what you have and are paid interest on the accrued amount? Then, you recieve $A(1+\frac{r}{2})((1+\frac{r}{2})$. One can continue in this vein until you get a more general formula for any m divisions of the year: $A(1+\frac{r}{m})^m$.

One may think of compounding as a bonus that the bank gives you for keeping your money there rather than withdrawing and reinvesting elsewhere. To appreciate the value of compounding, suppose that you invest \$100,000 in an account that offers 6 percent annual interest, compounded quarterly. How much would you earn in interest at the end of the year? How does that compare with the return on your investment without compounding? Without compounding, you would earn 6 percent on \$100,000, or \$6,000. Using the formula, with $m = 4$, you would earn \$6136.36, or slightly more than 2 percent more than without compounding.

What if the bank compounds continously? To isolate the effect of continous compounding, assume $A = 1$ and $r = 1$. The formula for the amount paid at the end of a year is, then, $\lim_{n\to\infty}(1+\frac{1}{n})^n = e$. The exponential function is the limit of the sequence

$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n.$$

With continuous compounding, the return on an investment of A dollars is Ae^r.

This simple formula makes it extremely easy to calculate and compare the return on different investments. Using the properties of limits we can consider the value of a multi-year investment. Specifically, from property 3, the return on a investment that is continuously compounded over t years is Ae^{rt}.

In the theory of investments it is important to be able to compare different investment opportunities made over varying time periods and weigh their value. One way to do this is to compute the Present Values of the investments. The compound interest formula gives us an extremely simple way to do so. If an investment pays B dollars after t years with an interest rate r, we would like to know the present value of that investment. We make this computation using the inverse: $Be^{-rt} = A$.

For example, a ten-year Treasury bond pays \$1,000 on a 3 percent interest rate. That means, in 10 years you can cash in the bond for \$1,000. How much is that bond worth in today's dollars? How much would you have to pay for that bond? Answer: \$740.82$(= 1000e^{-.03\times 10})$.

[Lecture 3 ended here]

D. Continuity

Limits allow us to give a more precise definition of continuity. A function $f(x)$ is said to be continuous if the limit as x approaches any point N in the domain of X is equal to the value of the function at that point, i.e., $f(N)$. Step functions are not continuous; polynomials are continuous. Some rational functions may not be as the denominator may tend to zero as it approaches a point.

Mathematical modeling often assumes continous functions, because analysis using continuous functions is much easier.

5. Differentiation

Measuring the effect of one variable on another variable is of central concern to social scientists and policy analysts. For example, investing in the stock market involves a certain amount of risk as the market as a whole fluctuates. Specific stocks also fluctuate at their own rate. Investors measure the risk of a stock relative to the market as the fluctuation in the individual stock's price as a function of the fluctuation in the market's average or overall price. If one makes a graph of the percentage change in a stock's price against the percentage change in the market's price, the relationship looks very linear. When the relationship between the indpedendent and dependent variables is a straight line, we can represent the effect as the slope of the line – the change in Y divided by the change in X.

Many problems require more complicated functions. Differentiation generalizes the concept of the slope beyond the linear case. The basic tool for analyzing continuous fucntions is differentiation, which is the calculation of the instantaneous rate of change of a continuous function. That is, if we change X an arbitrarily small amount, how much will y change?

A. Examples

i. Marginal returns

In a wide variety of problems we want to know how changing behavior somewhat will change outcomes. For example, how will adding more police on the street affect crime rates? It is usually too expensive to double the number of police, but how much will a modest increase, say a 5 percent increase, in the police force affect crime? Will crime change a little or a lot? Is the change in crime worth the cost of training and hiring the additional force?

Another example arises in your exercise. You are to use census data to measure the marginal returns to education. How much additional wages can you expect to earn for each additional year of education? Are the marginal returns decreasing, as standard economic theory of capital suggests, or increasing, as many recent arguments about rising inequality suggest?

You can measure how an additional year of education translates into higher wages by calculating the differential. That is, you are working with the function:

$$Y_t = \alpha_0 + \alpha_1 S_t + \alpha_2 S_t^2 + \alpha_3 X_t$$

The differential is the change in wages with respect to a unit change in one of the independent variables: $Y_t(S = s) - Y_t(S = s - 1)$. This is:

$$\Delta Y_t = \alpha_1(s - (s - 1)) + \alpha_2(s^2 - (s - 1)^2) = \alpha_1 - 2\alpha_2(s - 1)$$

A simple way to represent these phenomena is to model how differences in the independent variable translate into changes in the dependent variable. Since these variables are continuous, we need a way to measure rates of change for continuous functions.

ii. Risk Attitudes

Many problems imply that functions used to model behavior take a certain shape. An important distinction in studying shape of function is concavity and convexity. A convex function is "bowl" shaped; a concave function is "hill" shaped.

An important application is modeling decision making under uncertainty. A person is said to be risk neutral if he or she is indifferent between a "sure thing" and a gamble that has an expected or average value equal to the "sure thing." A person is risk averse if he or she prefers the sure thing to the gamble, and risk accepting if he or she prefers the gamble to the sure thing. A person's preferences can be represented by a "utility" function – a function that expresses the value that an individual places on money and consumption goods. The concavity of the utility function in terms of money captures the individual's attitudes.

A function is said to be concave if, for any two points, the value of the function at a weighted average of these two values is larger than the weighted average of the values of the function evaluated at the two points separately. If the opposite holds the function is convex.

Specifically, consider any two points, x_1 and x_2, in the domain of X. Let α be a number between 0 and 1, ie.,$0 < \alpha < 1$. The function $f(x)$ is concave if and only if $f(x = \alpha x_1 + (1 -$

$\alpha)x_2) > \alpha f(x_1) + (1-\alpha)f(x_2)$.

In decision-making under uncertainty, we think of α as the probability of the event that pays x_1 and $1-\alpha$ to be the probability of the event that pays x_2. The expected value of these two events is $\alpha x_1 + (1-\alpha)x_2$, which is sometimes called the certainty equivalent.

This mathematics is somewhat cumbersome, a more elegant description of risk attitudes can be arrived at using a measure of the rate of change of the rate of change of the utility function, i.e., the second derivative.

iii. Maximizing Behavior

Decision-making also involves making choices that give one the highest utility or the highest payoff. For example, in contests, such as military conflicts or political campaigns, a player chooses how much effort to expend to win a contest. Let X be the amount of effort spent to win the contest; let $P = X/(X+a)$ be the probability of winning the contest as a function of effort; and let the cost of effort be $c(X)$, say $c(X) = c_1 X$.

The payoff to the player is the probability of winning times the value of winning (say 1) minus the cost. So the player chooses X to maximize $U(X) = P(X) - c(X)$, that is to make $U(X)$ highest. The graph indicates that such a point exists

[Graph]

iv. Statistical Optimization.

Statistical modeling often begins with an objective function that summarizes the process that generated the data (the distribution of the data). The challenge for the statistician is the find the most efficient way to use the data to estimate the unknown parameters of the objective function. Typically, the objective function measures the variability of the data around the hypothesized model and

For example, regression analysis consists of finding the intercept and slope of the line that minimizes the total amount of "error," where the error is the squared deviation of the observed dependent variable from the predicted values. Let y_i be the observed value of the dependent variable and x_i be the observed value of the independent variable. The true relationship between these variables is $y_i = \alpha + \beta x_i + \epsilon$, where ϵ is "error." We wish to derive formulas for α and β in terms of things we observe, i.e., x_i and y_i. To do so, we choose α and β to minimize the total error:

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

This is a parabola in β and α. We wish to choose a and b that are at the lowest point of the bowl.

B. Definition of Derivative

The derivative is a continuous measure of the rate of change of a function. We define the derivative as the limiting value of the differential. We assume, therefore, that the function in question is continuous.

Consider a function of a single variable. Let h equal the difference between two values of X, i.e., $h = (x + h) - x$. The derivative is the slope of the line tangent to the curve at the point x.

$$\frac{dy}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Note: both the top and the bottom of this ratio tend to 0.

Analysis of specific functions will make clear the utility of the derivative.

Before proceeding to specific functions, we define a similar concept for a function of many variables. The partial derivative measures the rate of change in a function with respect to change in one variable holding constant all other variables. Let $y = f(x_1, x_2, ...x_n)$, then the partial derivative is defined as:

$$\frac{\partial y}{\partial x_j} = \lim_{h \to 0} \frac{f(x_1, x_2, ...x_{j-1}, x_j + h, ...x_n) - f(x_1, x_2, ...x_{j-1}, x_j, ...x_n)}{h}$$

Finally, we may take derivatives repeatedly to study more completely the shape of a function. For example, earlier we related risk attitudes to the rate of change of the rate of change of an individual's utility function. That is the derivative of the derivative, or the second-derivative. We denote this with the exponent 2, because we are essentially squaring the derivative:

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\frac{d}{dx}f(x)$$

For functions of many variables, we have a collection of second derivatives: $\frac{\partial y}{\partial x_j^2}$ and $\frac{\partial^2 y}{\partial x_j \partial x_k}$.

C. Specific Functions

i. Polynomials

First consider the straight line

$$Y = a_0 + a_1 X$$

Using the definition of the derivative:

$$\frac{dy}{dx} = \lim_{h\to 0}\frac{a_0 + a_1(x+h) - a_0 - a_1 x}{h} = \lim_{h\to 0}\frac{a_1 h}{h} = a_1$$

Second, consider a power function:

$$Y = a_1 X^n,$$

$n! = 1$. Using the definition of the derivative:

$$\frac{dy}{dx} = \lim_{h\to 0}\frac{a_1(x+h)^n - a_1 x^n}{h}$$

To analyze this we must expand the polynomial, using the binomial coefficient. Specifically, $(x+h)^n = x^n + \frac{n!}{(n-1)!1!}x^{n-1}h^1 + \frac{n!}{(n-2)!2!}x^{n-2}h^2 ... + h^n$. Now we can write the limit as

$$\frac{dy}{dx} = \lim_{h\to 0}\frac{1}{h}(x^n + \frac{n!}{(n-1)!1!}x^{n-1}h^1 + \frac{n!}{(n-2)!2!}x^{n-2}h^2 ... + h^n - x^n$$

The terms x^n cancel and we can divide each term by h to get:

$$\frac{dy}{dx} = \lim_{h\to 0}(\frac{n!}{(n-1)!1!}x^{n-1} + \frac{n!}{(n-2)!2!}x^{n-2}h ... + h^{n-1}$$

As h approaches 0 all of the terms in this expansion will approach 0 because all have a term h^m. Hence the limit equals the first value in the expression.

$$\frac{dy}{dx} = \left(\frac{n!}{(n-1)!1!}\right) x^{n-1} = nx^{n-1}$$

Finally, consider a general polynomial:

$$Y = a_0 + a_1 X^1 + a_2 X^2 + ...a_n X^n$$

Following the approach above one can show that:

$$\frac{dy}{dx} = a_1 + 2a_2 X + ...na_n X^{n-1}$$

Example. Marginal Return to Education.

In your homework you are to consider how an increase in schooling translates into increase wages. You may use the differencing formula above. You may also consider the formula as if the change is continuous:

$$Wages = \alpha_0 + \alpha_1 S + \alpha_2 S^2$$

$$\frac{dW}{dS} = \alpha_1 + 2\alpha_2 S$$

Returns to schooling are increasing if α_1 and α_2 are such that the first derivative is positive for all values of S is the domain (from 8 to 20 – note I do not presume that graduate education extends beyond 4 years).

We may further compute the second derivative of this function:

$$\frac{d^2W}{dS^2} = 2\alpha_2$$

The marginal returns to education are decreasing if $\alpha_2 < 0$.

ii. Logarithms

$$Y = log(X)$$

Assume natural logarithm, written $ln(X)$.

$$\frac{dY}{dX} = \lim_{h \to 0} \frac{ln(X+h) - ln(X)}{h}$$

Using the rule that the difference in logarithms is the log of a ratio, we can write this as

$$\frac{dY}{dX} = \lim_{h \to 0} \frac{1}{h} ln\left(\frac{X+h}{X}\right) = \lim_{h \to 0} \frac{1}{h} ln\left(1 + \frac{h}{X}\right)$$

Also, use the rule that exponents in logarithms equal coefficents:

$$\frac{dY}{dX} = \lim_{h \to 0} ln\left(1 + \frac{h}{X}\right)^{(1/h)}.$$

This looks very similar to the limit that defines the number e, except that h tends to 0. We can change the variable on the limit. Let $h = 1/m$. Then,

$$\frac{dY}{dX} = \lim_{m \to 0} ln\left(1 + \frac{1/X}{m}\right)^{m} = ln(e^{1/X}) = \frac{1}{X}.$$

Hence,

$$\frac{d}{dX} ln(X) = \frac{1}{X}$$

Note on Elasticities. This formula for the derivative of the logarithm is of considerable practical importance in statistics as researchers often measure variables in terms of logarithms (as mentioned in the last lecture). Typically, one might see a regression analysis along the following lines:

$$log(Y) = a + blog(X).$$

How do we interpret the coefficient in a regression analysis when the Y and X variables are measured in logarithms? To figure this out we take the derivatives of both sides with respect to X. Hence,

$$\frac{dlog(Y)}{dX} = \frac{d}{a + blog(X)} dX = b\frac{dlog(X)}{dX}$$

Analyzing both sides:

$$\frac{1}{Y}\frac{dY}{dX} = b\frac{1}{X}$$

Solving for b tells us how to interpret the coefficient in such a regression:

$$b = \frac{X}{Y}\frac{dY}{dX} = \frac{dY/Y}{dX/X}$$

The formulation on the right is the percentage change in Y divided by the percentage change in X. The coefficient b, then, represents the percentage change in Y for a one percent change in X. This concept is called an elasticity.

More generally, transforming variables using logarithms converts the scale from unit changes in a variable to percentage changes in a variable.

Example. Representation and Public Finances. Ansolabehere, Gerber, and Snyder (2002) estimate the relationship between log of Representation Per Capita and log of Public Funds Per Capita. They estimate a linear function of this log-log relationship, and the slope is .2. This means that a 100 percent increase in a county's state legislative representation corresponds to a 20 percent increase in the share of public funds received by a county.

iii. Exponentials

$$Y = e^X$$

$$\frac{dY}{dX} = e^X$$

D. Differentiation Rules

i. Product

$$Y = f(X)g(X)$$

$$\frac{dY}{dX} = \frac{df}{dX}g(X) + \frac{dg}{dx}f(X)$$

iii. Ratio

$$Y = f(X)/g(X)$$

$$\frac{dY}{dX} = \frac{g(X)(df/dX) - f(X)(dg/dX)}{[g(X)]^2}$$

Example. Contest.

$$U(X) = \frac{X}{X+A} - cX$$

$$\frac{dU}{dX} = \frac{1(X+A) - 1X}{[X+A]^2} - c = \frac{A}{[X+A]^2} - c$$

iii. Chain

$$Y = g(f(X))$$

$$\frac{dY}{dX} = \frac{g(f)}{df}\frac{df}{dX}$$

Example. The Shape of the Normal Curve.

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

$$\frac{df}{dz}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}(-z),$$

so this function is increasing for negative values of z and decreasing for positive values of z. Using the product and the chain rules:

$$\frac{d^2f}{dz^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}(-z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}(-1) + \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}(-z)^2 = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}(z^2-1),$$

so the second derivative is negative if z lies between -1 and +1, i.e., the function is concave, and the second derivative is positive if z lies outside the interval -1 and +1.

Example. Sums of Squared Errors.

We wish to measure the rate of change in total error with respect to *both* α and β. This involves partial differentiation. We implement this by considering the change in one of the variables, assuming the other equals a constant.

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

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

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

iv. Implicit Function

$$h(Y, X) = 0$$

$$\frac{\partial h}{\partial Y} \frac{dY}{dX} + \frac{\partial h}{\partial X} = 0$$

or

$$\frac{dY}{dX} = -\frac{\partial h / \partial X}{\partial h / \partial Y}$$

Example. Cube Law.

A more general formula for the relationship between seats and votes is

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

What is dS/dV? The derivative of the lefthand side with respect to V is $\frac{1}{[1-S]^2} \frac{dS}{dV}$. The derivative of the righthand side is: $ab(\frac{V}{1-V})^{b-1} \frac{1}{[1-V]^2}$. Isolating the derivative:

$$\frac{dS}{dV} = (1-S)^2 ab(\frac{V}{1-V})^b \frac{1-V}{V} \frac{1}{[1-V]^2} = b(1-S)^2 \frac{S}{(1-S)} \frac{1}{V(1-V)} = b\frac{S(1-S)}{V(1-V)}$$

E. Maximization and Minimization

A number of the problems that are of interest involve finding the value of X that makes Y smallest or largest. Maximizing behavior and statistical optimization both involve finding such "extreme values." Differentiation offers a natural way to find such values.

At a maximum or a minimum of a continuous function, the instantaneous rate of change in the function equals 0. At a maximum it must also be the case that this value is the top of a hill, so the function is concave at this point. At a minimum it must be the case that this value is the top of a hill, so the function is concave at this point.

So a local extreme value can be found by solving the equation or first order condition:

$$\frac{dY}{dX} = 0$$

One may verify that the function is concave or convex by determining the sign of the second derivative, the second order condition. When Y depends on many variables, the first order condition holds that all partial first derivatives equal 0. The second order condition is somewhat more complicated and we leave that to a more advanced course

To develop this concept, let us solve the problems that we have developed during this unit.

Example. Least Squares Solution to the Statistical Error Function.

To find the values of α and β that minimize total squared error set the first derivatives equal to zero and solve. Let $\hat{\alpha}$ and $\hat{\beta}$ be the values of the parameters that solve the problem. The first order conditions, sometimes called "normal equations," are:

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

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

To solve the first equation, begin by carrying through the sum:

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

Divide by -2 and isolate $\hat{\alpha}$:$n\hat{\alpha}\sum_{i=1}^{n} -2y_i + \sum_{i=1}^{n} \hat{\beta}x_i$, so

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

Algebraic manipulation of the second normal equation reveals that

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

which is the covariance between X and Y (a measure of the "rise") divided by the variance of X(a measure of the "run"). Note, estimates of both parameters can now be expressed as functions of the data, observed values of X and Y.

Example. Maximizing Behavior.

What is the optimal level of effort to devote to winning a costly contest? A player chooses effort level X, given the fixed effort of others, A, and the cost function cX. Denote the optimal effort level as X^*.

$$U(X) = \frac{X}{X+A} - cX$$

The first order condition is

$$\frac{dU}{dX} = \frac{A}{[X^* + A]^2} - c = 0$$

To isolate X^*, put c on the righthand side, multiply both sides of the equation by $(X^* + A)^2$ and take square roots. Then,

$$X^* = \sqrt{\frac{A}{c}} - A$$

If A is large enough this equation is 0, which is out of bounds, so we also have $X^* = 0$ if $A > 1/c$.

To verify that this is a maximizing strategy take the second derivative of the utility function:

$$\frac{d^2U}{dX^2} = -\frac{2}{[X^* + A]^3} < 0,$$

for $X > -A$.

Example. Nash Equilibrium.

What if A is a player too? Suppose both players will choose the level of effort that maximizes their utilities. How will they strategize about this problem. Each will assume that the other is doing the best he or she can. So, player X will maximize his or her own utility assuming that A plays an optimal strategy A^*, and A will assume that X is playing an optimal strategy. Since the game is to increase the probability of winning, the probability that X wins equals 1 minus the probability that A wins, so the two players' utility functions are:

$$U(X; A^*) = \frac{X}{X + A^*} - cX$$
$$W(A; X^*) = (1 - \frac{X^*}{X^* + A}) - cA$$

The equilibrium is defined by the solution to both of these problems simultaneously. That is, both equations must satisfy their first order conditions:

$$\frac{\partial U}{\partial X} = \frac{A^*}{[X^* + A^*]^2} - c = 0$$

$$\frac{\partial W}{\partial A} = \frac{X^*}{[X^* + A^*]^2} - c = 0$$

So $X^* = A^* = \frac{1}{2c}$. Notice how different the solution to a game in which two people simultaneously make a decision as opposed to a situation where one must take the other player as given.

Comparative Statics. The solution to a problem, such as the optimal effort level in a contest, is a function of parameters. We may further analyze how the predicted outcome changes as the parameters vary. Such calculations are called comparative statics – predicted variation in observed behavior as important factors vary. For example, in the decision-theoretic version of the contest, optimal effort is a function of the marginal cost and the parameter A, which one might think of as the size of the incumbent army or candidate. The solution offers a specific functional form: $X^* = \sqrt{A/c} - A$. The comparative statics are:

$$\frac{\partial X^*}{\partial c} = -\frac{1}{2}A^{-1/2}c^{-3/2}$$

$$\frac{\partial X^*}{\partial A} = \frac{1}{2}\sqrt{\frac{c}{A}} - 1$$

The first comparative static means that the higher the marginal cost the lower the total effort. The second comparative static has an ambiguous sign. It is positive if $A < c/4$, an admittedly small number.

F. Taylor's Approximation.

A final use of derivatives is that we can approximate any function using a polynomial. This result is commonly used in statistical analyses, as we often use the first few powers of a function to approximate a function. The result stems from Taylor's Polynomial Expansion.

$$f(x) \approx \frac{x^0}{0!}f(0) + \frac{x^1}{1!}\frac{df(0)}{dx} + \frac{(x)^2}{2!}\frac{d^2f(0)}{dx^2} + \frac{x^3}{3!}\frac{d^3f(0)}{dx^3} + ...$$

More generally, we can make the approximation around any point μ.

$$f(x) \approx \frac{(x-\mu)^0}{0!}f(\mu) + \frac{(x-\mu)^1}{1!}\frac{df(\mu)}{dx} + \frac{(x-\mu)^2}{2!}\frac{d^2f(\mu)}{dx^2} + \frac{(x-\mu)^3}{3!}\frac{d^3f(\mu)}{dx^3} + ...$$

We can choose any degree of approximation we desire by incorporating higher orders of the expansion.

Examples. $f(x) = e^x$.

$$f(x) \approx e^0 + xe^0 + \frac{x^2}{2!}e^0 + \frac{x^3}{3!}e^0 + \frac{x^4}{4!}e^0.... = \sum_{i=0}^{\infty}\frac{x^i}{i!}$$

This was one of our definitions of the function e^x.

$f(x) = log(x)$.

$$f(x) \approx log(\mu) + \frac{(x-\mu)}{x} - \frac{(x-\mu)^2}{2x^2} + 2\frac{(x-\mu)^3}{6x^3} - 6\frac{(x-\mu)^4}{24x^4}...$$

Let $\mu = 1$, then

$$f(x) \approx \frac{(x-1)}{x} - \frac{(x-1)^2}{2x^2} + 2\frac{(x-1)^3}{6x^3} - 6\frac{(x-1)^4}{24x^4}...$$

Exercise Use Taylor's theorem to approximate the normal curve $e^{-.5x^2}$

6. Summation and Integration

Summation and integration are central to probability and statistics. There are two important reasons. (1) Most statistics are averages, so we will need to understand how to sum variables. (2) Once we have estimated parameters of interest using statistics we will make inferences and predictions using our estimates, but our estimates come with some degree of uncertainty. Any inferences or predictions we make, then, are about not a single point but a range or set of observations. We will use probability to make statements about our degree of confidence. We will have to sum the total amount of probability associated with a given set or range in order to make statements about our degree of confidence in our estimates. [This is a tricky issue. It is a matter of human nature that we are always more confident in hypotheses then we perhaps should be.]

We can think of the inference problem two ways. First, it is a sum: we are summing all of the probability associated with all of the possible points in a set. Second, it is an area. The probability curve describes the likelihood of every point along a line (or on a plane), and the hypothesis is a prediction about a set along the line or on a plane (ie., the points between a and b). Inference involves calculating the total area under the curve associated with the set predicted by the hypothesis.

Integration is also commonly used in formal modeling, especially when there is a dynamic relationship or when uncertainty is involved. Let's consider a simple uncertainty problem. Suppose that X has an equal chance of occuring anywhere along the interval 0 to 1, but cannot occur outside that interval. What value of X do we expect, i.e., what is the certainty equivalent?

A. Notation.

For discrete numbers, $x_1, x_2, x_3 ... x_n$, we define the sum as:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + ... + x_n$$

Similarly, for any function of X we can calculate the sum: $\sum_i g(x_i)$. We also may have variables summed over different indexes. In statistics such sums appear when there are different levels to data, such as individuals i and counties j.

$$\sum_{j=1}^{m} \sum_{i=1}^{n} x_{ij} = (x_{11} + x_{21} + x_{31} + ... + x_{n1}) + ... + (x_{1m} + x_{2m} + x_{3m} + ... + x_{nm})$$

The summation operator has the properties of any linear function. Several appear repeatedly in statistics, especially the features of sums of constants, linear functions and quadratics of sums:

$$\sum_{i=1}^{n} x_i = (a + bx_1) + (a + bx_2) + (a + bx_3) + ... + (a + bx_n) = na + b \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} (x_i + y_i) \right)^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 + 2 \sum_{i>j} x_i y_j$$

When X is a continuous variable, we perform a conceptually similar operation, but as with differentiation and differencing care needs to be taken in defining how the summation occurs. The notation for the sum of a continuous variable and a function of a continuous variable, $f(x)$, is defined:

$$\int f(x) dx$$

As a sum this has the properties of other sums, e.g., $\int (a + bf(x)) dx = a \int dx + b \int f(x) dx$.

B. Integral defined.

The integral may be thought of three ways. (1) It is the sum of the values of a continuous function. (2) It is the area under a curve. (3) It is the inverse of the derivative. We have already introduced the notation as the sum of a continuous function. The second definition will allow us to construct the integral as a limit. The third definition is of great utility as it tells us a general strategy for solving integration problems: imagine the functions for which the function inside the integral is the derivative. A function $F(x)$ such that $dF/dx = f(x)$ is called a primitive of $f(x)$.

Let's define the integral more carefully using the second concept – the area under the curve. We would like to calculate the area under the curve $f(x)$ between two points, a and b. To do this let's divide the problem up into many discrete calculations and then take the limit.

Divide the x-axis into n (equal or unequal) intervals. Let the points $x_1, ...x_{n-1}$ be the points that divide the line. The end points are $x_0 = a$ and $x_n = b$. Let the difference between successive points on the line be $\delta x_j = x_{j+1} - x_j$. Within any interval choose an arbitrary point, say μ_j. For any one of segment, calculate the height of the function at μ_j. We can calculate the approximate area under the curve in a given segment of X as the area of the rectangle with height $f(\mu_j)$ and base δx_j. This area is $f(\mu_j)\delta x_j$. To calculate the approximate area under the curve sum all of the rectangles:

$$F_n = \sum_{j=1}^{n} f(\mu_j)\delta x_j = f(\mu_1)(x_1 - x_0) + f(\mu_2)(x_2 - x_1)...f(\mu_n)(x_n - x_{n-1})$$

The badness of the approximation is the sum of the "triangular"-shaped curves left over from each of the rectangles. [GRAPH]

The integral is defined as the limit as the number of divisions (n) increases without bound and the width of the largest interval (δx_j) shrinks to 0.

The integral over a specified region (say a to b) is called the <u>d</u>efinite integral. When boundaries are variable or when no boundaries are specified, the integral is called the <u>i</u>ndefinite integral. It is important to note that different definite integrals of the same function differ only by a constant.

More specifically, we denote the indefinite integral as the integral from a to x, where a is a fixed lower bound, but the upperbound x varies.

Two important theorems from calculus are enlightening for probability theory and statistics. We'll state them here.

Fundamental Theorem

The difference of two primitives $F_1(x)$ and $F_2(x)$ of the same function $f(x)$ is always a constant: $F_1 - F_2 = k$. From any one primitive we can obtain all other primitives with a suitable choice of c. So,

$$F(x) = c + \int_a^x f(u)du$$

Likewise, for any function $f(x)$ the primatives are equivalent up to a constant. This provides us with our third definition of the integral and links integration to differentiation.

From the fundamental theorem it follows that the value of a definite integral can be calculated as the value of the primitive function evaluated at the upper limit b minus the value of the primative function evaluated at the lower limit a. That is

$$\int_a^b f(x)dx = F(b) + c - (F(a) + c) = F(b) - F(a)$$

Note: the constant term in the indefinite integral cancels giving a definite solution to the problem.

Mean Value Theorem

Let m be the smallest value of the function $f(x)$ and M be the largest value of the function $f(x)$ on the interval a, b. Then there exists some point μ between m and M such that

$$\int_a^b f(x)dx = \mu(b - a)$$

C. Calculation of Integrals.

Using the area definition and the fundamental theorem offer two ways to calculate integrals.

Let us consider a very simple function: $f(x) = k$ if $a \leq x \leq b$, and 0 otherwise. Viewed as an area, the integral is the area of a rectangle, the base times the height, i.e., $k(b-a)$. Viewed as the inverse of the derivative (the fundamental theorem), the integral is the indefinite function for which a constant term k is the derivative. That is $F(x) = kx + c$. To calculate the definite integral we calculate the difference between the primative function at the upper bound and at the lower bound, $F(b) - F(a) = kb + c - ka - c = k(b - a)$.

Consider polynomials.

A linear $f(x)$ yields a quadratic $F(x)$; a quadratic $f(x)$ yields a cubic $F(x)$ and so forth.

$$\int xdx = \frac{1}{2}x^2 + c$$

$$\int x^2dx = \frac{1}{3}x^3 + c$$

$$\int x^ndx = \frac{1}{n+1}x^{n+1} + c$$

Consider logarithms and exponentials.

You can verify using the fundamental theorem that

$$\int log(x) = xlog(x) + c$$

$$\int e^x = e^x + c$$

More complicated functions require more intricate integration. Some integrals cannot be solved in "closed form" – we require a computer. One such integral is $\int e^{x^2}dx$. Many problems can be solved with creative use of the differentiation rules. A common trick called u-substitution involves the chain rule. Imagine that the function is such that $f(x) = (dg/dx)(df(g(x))/dg)$, then $\int f(x)dx = F(g(x)) + c$. For example, consider $f(x) = xe^{x^2}$. Let $u = g(x) = x^2$. Then $du/dx = 2x$, so $du = 2xdx$. Because there isn't a 2 in the formula divide by 2 and multiply by 2. We can then substitute in for u to get:

$$\int xe^{x^2}dx = \int \frac{1}{2}e^udu = \frac{1}{2}e^u + c = \frac{1}{2}e^{x^2}$$

Example. In the analysis of the relationship of the distribution of public expenditures to representation, Ansolabehere, Gerber, and Snyder (2002), regress share of public expenditures in logarithms on share of representation in logarithms:

$$log(S) = \alpha + \beta log(R)$$

If we were to equalize representation how much money would underrepresented counties, i.e., those with R less than 1, receive? To calculate the predicted values first convert back into the original units: $S_t = e^{\alpha} R_t^{\beta}$. The variable R ranges from 0 to a large value, but the smallest observed value is about .1. Second, to calculate how much underrepresented counties were "due" we calculate the integral from R = .1 to R = 1.

$$\int_0^1 SdR = \int_0^1 e^{\alpha} R^{\beta} dR = \frac{e^{\alpha}}{\beta + 1} R^{\beta+1}|_{.1}^{1} = e^{\alpha}[1 - \frac{1}{\beta + 1}.1^{\beta+1}]$$