A REVIEW OF SOME ALGEBRA, CALCULUS, AND STATISTICS NECESSARY FOR ITEM RESPONSE THEORY

John H. Neel

Department of Educational Foundations
Georgia State University

For complete understanding of Item Response Theory, you should probably have a good knowledge of a number of topics generally covered in calculus and mathematical statistics courses. There are those who seem to do quite well in IRT without knowledge of these topics but I feel that it must be quite difficult for them at some times.

The intention in this writing is to give you enough of these topics to remove some of the mystery from IRT models. These topics will be review for some of you and introduction for others. It is almost impossible to do justice to the topics in any short time period. My hope is that when the topics come up in the study of IRT, you will feel comfortable with the concepts.

Assumptions made for this writing are that you have had an algebra course some time in your life, and that you have had at least an introductory course in statistics. If you do all the exercises, you should feel more comfortable reading item response theory than you would if you had not done them.

ALGEBRA

Everyone probably remembers most of the algebra necessary for item response theory. This is because you have already been in enough statistics courses to keep at least some of your knowledge of algebra reasonably current. Some infrequently used areas are also essential: exponents, logarithms, and e. These are discussed here.

Exponents

The following rules regarding exponents are usually covered in a high school algebra I class.

1. $X^{-Y} = \dfrac{1}{X^{Y}}$ and $X^{Y} = \dfrac{1}{X^{-Y}}$

2. $P^{X} \cdot P^{Y} = P^{X+Y}$

3. $P^{X}/P^{Y} = P^{X-Y}$

4. $P^{X} \cdot Q^{X} = (PQ)^{X}$

There isn't much to do for these but memorize them. Some explanation of why they mean what they do will be given in class. The best approach to memorizing the rules is to work lots of simple problems like those in the set of exercises following these first few sections.

Logarithms

One of the first things to learn about logarithms is that there are different systems of logarithms which have different bases. In general, two systems of logarithms cannot be mixed without conversion from one to the other. There are formulas for converting from one base to another, but these are not particularly useful to us because we will only work in one base. When logarithms are being used, the user must know what base is being used and not mix bases. This initial discussion will assume a base of 10. When 10 is used as the base we use the term 'common logarithms'.

The logarithm of the number, x, is a number, m, such that

$$10^{m} = x .$$

In other words, the logarithm of x is the number, m, which when 10 is raised to the m power gives us x. The logarithm of 10 is 1 because

$$10^1 = 10.$$

The logarithm of 100 is 2 because

$$10^2 = 100.$$

The logarithm of 100000 is 5.

What is the logarithm of 1000000000000000?

When base 10 logarithms are used it is customary to indicate this by either writing

$$\log_{10}(X) \quad \text{or simply}$$

$$\log(X).$$

We can thus write:

$$\log_{10}(100) = 2, \quad \text{and}$$

$$\log(1000) = 3.$$

Since logarithms are exponents, they have the properties of exponents. Rule 2 of the section on exponents says

$$P^x \cdot P^y = P^{x+y}$$

This fact can be used to do multiplication by taking the logarithms of the numbers to be multiplied, adding the logarithms, and then raising the base to the sum of the logarithms. We are using rule 2 from the section on exponents. For example if the product

$$100 \cdot 1000$$

is needed, one can notice that

$$\log(100) = 2, \quad \text{and}$$

$$\log(1000) = 3.$$

Thus, $100 \cdot 1000 = 10^{2+3} = 10^5 = 100000.$

Similarly

$$\log(57) = 1.7559$$
$$\log(18) = 1.2553$$

thus $57 \cdot 18 = 10^{1.7559 + 1.2553} = 10^{3.0112} \approx 1026.1$

Other bases rather than 10 can be used for logarithms. In some applications, base 2 or base 16 is convenient. In statistics and measurement, as in much of natural science, the constant

$$e = 2.718282...$$

is convenient to use as a base[1]. The logarithms using e as a base are often referred to as 'natural logarithms' or 'Naperian logarithms' after the Scottish mathematician John Napier who is credited with the discovery of logarithms. When base e is used, the customary notation is

$$\log_e(x) \quad \text{or}$$

$$\ln(X) \ .$$

This latter notation, ln(x), is the notation preferred by most authors and is the notation to be used here.

In most measurement or statistical applications the natural logarithm is used because many formulas and theories take their simplest form using natural logarithms.

Logarithms show up in statistics and are often useful by being another way to do multiplication. Sometimes taking logarithms can greatly simplify theory. This usually happens as a result of moving from multiplication to addition. The theory is then made more pliant because addition is simpler than multiplication.

The natural logarithm base, e

The natural logarithm base, e, is a constant like π which occurs frequently in mathematics, statistics, and the natural sciences. Also, e frequently appears in scientific models when logarithms are not being used. The number may be found by taking the following limit as n approaches infinity.

$$\lim_{n \to \infty} (1 + 1/n)^n$$

For n from 1 to 10 the term $(1 + 1/n)^n$ evaluates as given in Table 1.

---

1. The next section discusses e, the base of natural logarithms.

Table 1

First ten terms of a sequence with e as the limit.

n	$(1 + 1/n)^n$
1	2.
2	2.25
3	2.37037
4	2.44140
5	2.48832
6	2.52162
7	2.54650
8	2.56578
9	2.58117
10	2.59374

As you can see in Table 1, the value of the expression gets larger as n gets larger. However it is a case of "diminishing returns" and there is a limiting value to which the expression will come arbitrarily close and never exceed. The limiting value (to 10 decimal places) of the series is 2.7182818285 and is usually written as e.

This number appears, among other places, in the equation for the normal curve:

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

This equation gives the height of the normal curve at x when the curve has mean μ and standard deviation σ. If the curve is a standard normal curve, mean 0 and standard deviation 1, we write f(z) instead of f(x) to indicate this and the function reduces to:

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

This is the version which is tabled in normal curve area tables.

Exercise Set 1

1. Use the exp, inverse ln, or similar function on a hand calculator to find the value of e. Give directions for doing this on your calculator. (Remember, $e^1 = e$.)

2. Make a table similar to Table 1 by using a hand calculator to evaluate the expression $(1 + 1/n)^n$ for values of n from 10 to 100 in increments of 10, from 100 to 1000 in increments of 100, and from 1000 to 10000 in increments of 1000. (If you know a programming language, this is probably even easier to program than to do on a hand calculator.) How close to the value given (to 10 decimal places) is the final answer?

3. Use the ln function on a hand calculator to find the natural logarithm of 2.7182818285 . Write directions for doing this.

4. Find the natural logarithms of the following numbers with a calculator.

   a) 2.
   b) 7.389056099
   c) 10
   d) 100
   e) 374
   f) 32.1889
   g) 5.36668

5. Use the natural logarithm functions on a hand calculator to find the following products (Find the logs, add them, raise e to that power.). Check your work by regular multiplication.

   a) $87 \cdot 19 =$

   b) $126 \cdot 43 =$

   c) $.00234 \cdot 145 =$

   d) $135 \cdot 35.26 =$

   e) $.0145 \cdot .00035 =$

6. Simplify the following expressions:

   a) $X^2 \cdot X^4 =$

   b) $X^4/X^2 =$

   c) $X^2/X^4 =$

   d) $X^3 \cdot Y^3 =$

   e) $X^Y \cdot X^Z =$

   f) $Y^5 \cdot Y^8 =$

   g) $X^8/X^{12} =$

   h) $X^{2.5} \cdot X =$

7. Calculate the height of the standard normal curve at z = 0.

Calculus

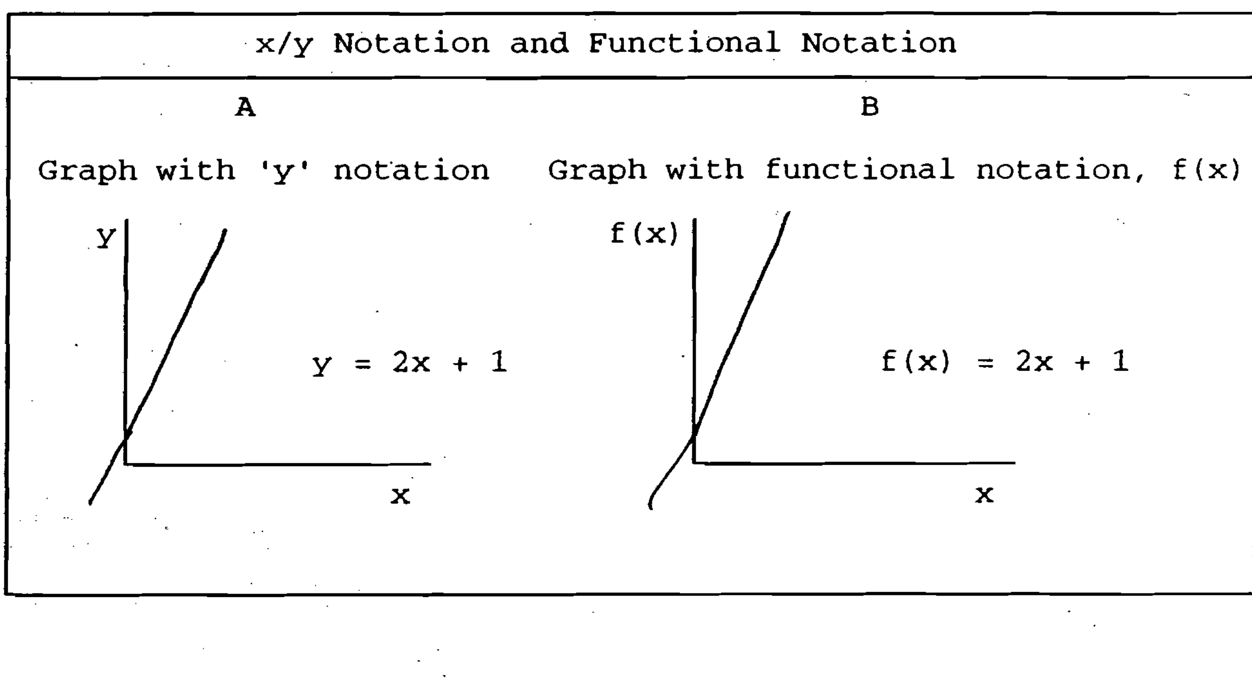
In most mathematics departments, calculus is taught 1 hour per day, 5 days a week, for an academic year. This fact causes me to hesitate as I write this. Just how much can you be reasonably expected to know and understand after the short instruction you will have here? What misconceptions will you have after I omit so much in order to get at the little that we can use in item response theory? After posing these questions, I took my hands off the keyboard and thought about this for some time. I could not come to an answer. Let's take that as a warning, dig in, and see how well we can do.

Two main branches of calculus are differential calculus and integral calculus. In differential calculus the slope of a tangent line to a curve is found. In integral calculus, areas under curves are found. Differential calculus is studied first since, as we will see, it logically precedes integral calculus in the theory. Both branches of calculus consider continuous functions, so we start with functional notation and continuity.

Functions

You probably remember graphing equations in introductory statistics classes. Figure 1 A graphs the linear equation $y = 2x + 1$.

Figure 1

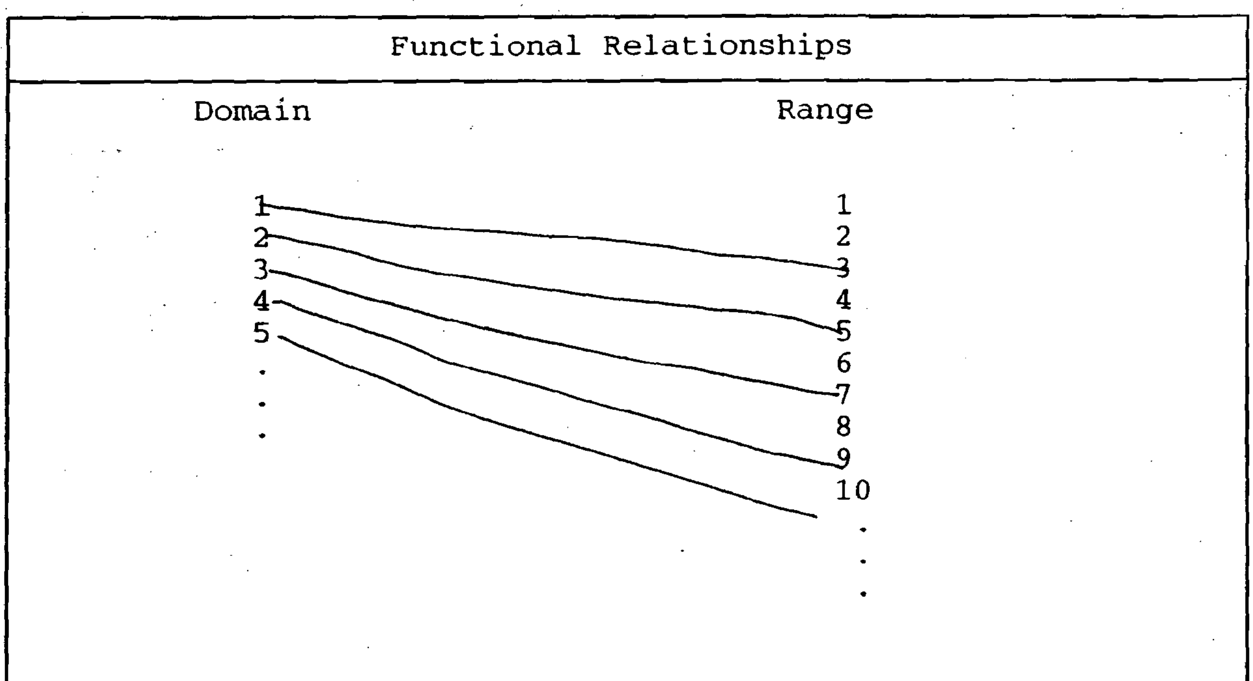


in Figure 1 A, you see y plotted on the vertical axis, the ordinate, and x plotted on the horizontal axis, the abscissa. it is common to refer to y as a function of x. This is in the sense that the value of y is dependent on the value of x. Another common terminology is to refer to the variable on the horizontal axis as the independent variable and the variable on the vertical axis as the dependent variable.

We usually indicate the functional relationship as f(x). Figure 1 A could be relabeled as in Figure 1 B with f(x) replacing y. Here, y and f(x) are synonyms. The f(x) notation is used because it makes it evident that a functional relationship exists and because it has certain ease of use advantages which will become apparent as we use the notation.

Functions are very formal and important relations in mathematics. Functions are assignments or rules for pairing of elements in one set with elements in another set. The first set is called the domain of the function, the second set is called the range of the function. The function is said to be a mapping from one set to the other set. This is illustrated for the function $f(x) = 2x + 1$ in Figure 2.

Figure 2



The rule for assigning an f(x) value to a given x has only the restrictions that

a) each x in the domain must be assigned a value in the range.
b) each x in the domain can only be assigned one value in the second set.

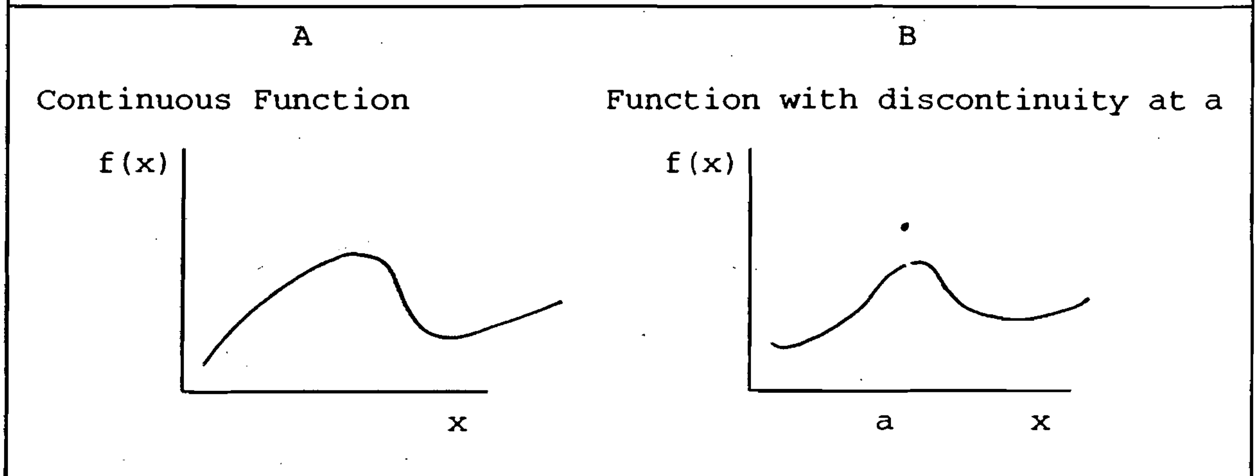
It is usually important that the domain and range of a function be well defined. The definitions of the domain and range are frequently implicit in the definition of the function and are thus not always discussed. For the function $y = 2x + 1$, both the domain and the range are the real numbers. Frequently, the domain of the function is restricted. The function $f(x) = \sqrt{x}$ has the domain of all nonnegative real numbers which restricts the domain of the square root function to zero and positive real numbers.

Continuity

Much of what can be done in calculus and other areas of higher mathematics is dependent on having continuous functions. Intuitively, a continuous function is one in which there are no holes. Consider an arbitrary smooth curve as illustrated in Figure 3 A. If we remove one point from the curve, and move it up (or down) the function is no longer continuous due to the gap at that point. This is shown in Figure 3 B.

Figure 3

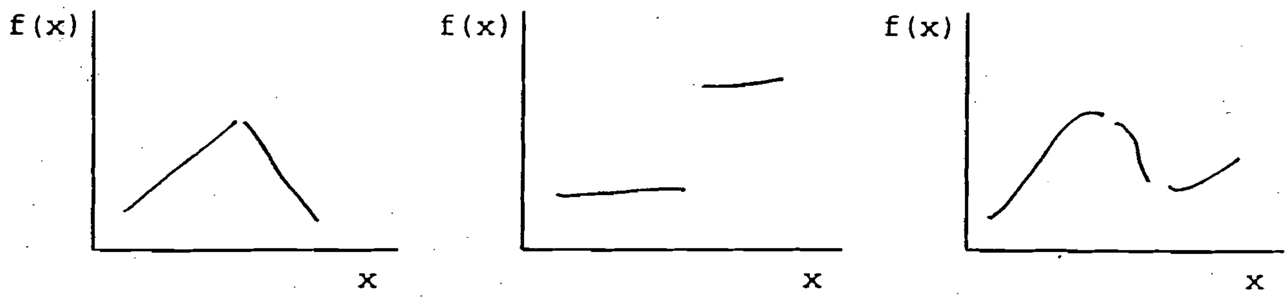
Graphs of Continuous and Discontinuous Functions



Discontinuous functions cause problems when working at or near the point of discontinuity. The "gap" in the function causes problems with finding limits of the function of x. Considering limiting values is frequently what is done in calculus. Thus the restriction to continuous functions. Figure 4 shows some additional functions with a point of discontinuity .

Figure 4

Graphs of functions with discontinuities



Differential Calculus

Consider Figure 5. In that figure, an arbitrary functions has been drawn. The first has a maximum value. The second has a minimum value. Functions which might look like this are

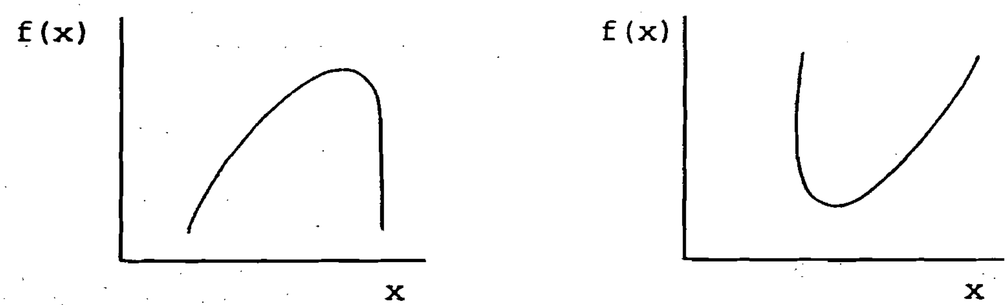
$$f(x) = -x^2 + 10x - 25$$

and

$$f(x) = x^2 + 5 .$$

Figure 5

A	B
Function with a maximum Value	Function with a minimum value

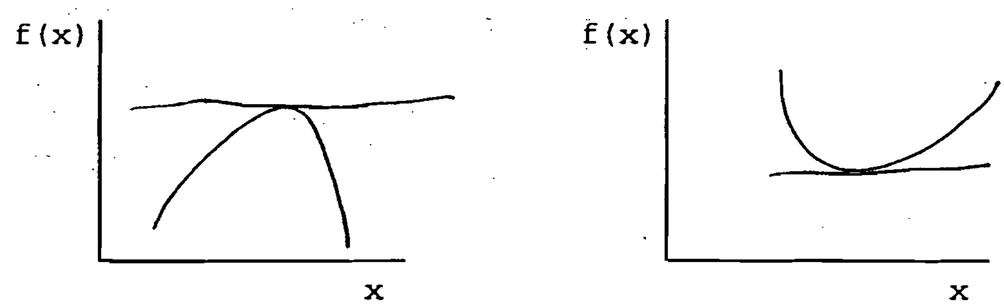


Simple functions similar to this can represent some very important concepts. They could represent the height of an artillery shell x minutes after leaving the cannon; the yield of a chemical process as the temperature, x, is raised; or (nota bena) as in Figure 5 B the sum of the squared differences between actual and predicted values in a regression problem. Someone investigating these functions might well want to know where the maximum or minimum value occurs. For what x is f(x) the greatest, or least, value? At what time does the artillery shell reach its highest point? At what temperature does the chemical process result in the maximum yield? Where does the minimum sum of the squared error terms occur?

You might suggest that all one needs to do is draw the graph and find the maximum by looking at the graph. That works for some simple cases, but as the functions grow more complex it turns out to not be very satisfactory. An analytic approach is necessary, one which yields an equation or formula which provides the answer. Figure 6 illustrates the approach. Notice that the tangent lines to the curve at the maximum and minimum are horizontal with a slope of zero.

Figure 6

Tangent Lines to functions



Tangent lines are drawn to the function. Then the slope of the tangent line is found. Next, reason that when the slope is zero, the tangent has been drawn at a maximum or minimum value. Solution found. Well, not quite found. First problem: is it a maximum or a minimum? Next problem: some functions have more than one peak, won't there be a tangent line with a zero slope at more than one point? The time it takes to answer to these and other questions is one of the reasons that calculus is studied for a year.

One complication we do need to clear up is that of a relative maximum or minimum. It is possible for a function to have more than one point where the function has a peak or trough. In relation to the "nearby" values of the function, the top or bottom of the peak or trough is the largest or smallest value. It is thus referred to as a relative maximum or minimum. Many functions have relative maxima or minima. In fact, a function can have a relative maximum or minima without having an absolute maximum or minimum. Maximum and minimum values are often referred to as the extreme values of the function or extrema.

The slope of the tangent line to the function is called the derivative. We find the derivative, see where it is zero, and that is where the maximum or minimum is. All of this can be done without actually graphing the functions. The graphs help us visualize what is happening, but after we understand, they are not necessary. Now it turns out that finding the derivative depends on just what kind of function we are considering. Sometimes the process is straight forward, and sometimes it is more complex. We are going to consider several of the straightforward rules for finding derivatives:

- the derivative of a constant,
- the power rule,
- the derivative of a sum (difference), and
- the chain rule.

To save time, I am not going to explain how these rules are derived, only how to use them. If you are interested in mastering these topics, there is probably no better way than to take the calculus series in the mathematics department.

<u>Some notation</u>

Something as important as a derivative has to have its own notation. There are several notational systems for derivatives. The one we will use is based on the functional notation. If the function of x is f(x) then we indicate the derivative by writing

$$f'(x)$$

or sometimes simply

$$f'.$$

This is read as "the derivative of the function f" or simply as "the derivative".[2]

Derivative of a constant

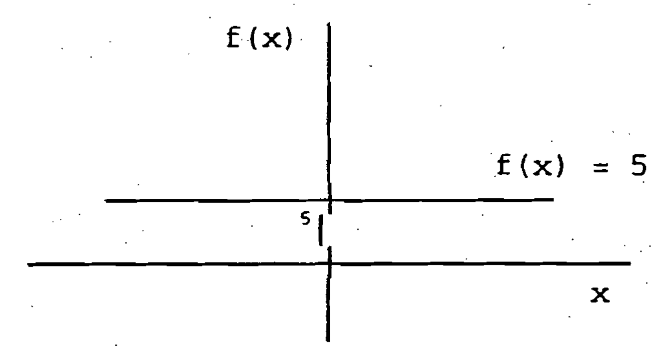
This is the simplest derivative. The derivative of any constant is 0. We can now take the derivative of any constant function as you can see here:

Function	Derivative
$f(x) = 5$	$f'(x) = 0$
$f(x) = 12.4$	$f'(x) = 0$
$f(x) = 290$	$f'(x) = 0$
$f(x) = 376.15$	$f'(x) = 0$
$f(x) = k$	$f'(x) = 0$

Let us consider the graph of the first function which is plotted in Figure 7. The graph of a constant function is simply a horizontal line through the y axis at the point equal to the constant. The tangent line to the curve is also a horizontal line at every point of the function. Thus, the derivative ends up being 0 at every value of x and every f(x) is a maximum value; or, if you want to consider it another way, every f(x) is a minimum value.

Figure 7

Graph of a constant function



[2] There are other common notations for derivatives. One of the most common is

$$\frac{dy}{dx}.$$

Power rule

With the power rule we learn how to take the derivative of a function which is an integral power of x. If f(x) can be expressed in the general form

$$f(x) = ax^n$$

where n is an integer, then the derivative is given by

$$f'(x) = nax^{n-1}.$$

To take the derivative you multiply the function by the power and reduce the power by one. Some examples:

Function	Derivative
$f(x) = x$	$f'(x) = 1x^{1-1} = x^0 = 1$
$f(x) = x^2$	$f'(x) = 2x^{2-1} = 2x^1 = 2x$
$f(x) = x^5$	$f'(x) = 5x^4$
$f(x) = 8x^9$	$f'(x) = 72x^8$

Derivative of a sum

The derivative of a sum, or difference, is the sum, or difference, of the derivatives. You simply take the derivatives of the parts and add, or subtract. This is also straight forward:

Function	Derivative
$f(x) = x + x^2$	$f'(x) = {}^1 + 2x$
$f(x) = 2x^3 - x^7$	$f'(x) = 6x^2 - 7x^6$
$f(x) = 4x^3 - 9$	$f'(x) = 12x^2 - 0 = 12x^2$
$f(x) = 2x^3 + x^2 - x$	$f'(x) = 6x^2 + 2x - 1$

Chain rule

The chain rule is useful when you have a complex function which can be simplified by considering it to be a function of a function. Consider the function

$$f(x) = (x + 4)^2.$$

This is not a particularly complicated function, but it does not fit any of the patterns we have so far. It is not a constant and it is not in the proper form to be a power function. If we define two new functions, the chain rule can be applied. The two new functions are:

$$g(u) = u^2$$

and

$$h(x) = x + 4.$$

What you should be able to see this far is that we can write $f(x)$ in terms of $g(x)$ and $h(x)$ where $u = h(x) = x + 4$.

$$f(x) = g(u) = g(\ h(x)\ ) = g(\ x + 4\ ) = (x + 4)^2$$

Here $f(x)$ is shown to be the composite of outer and inner functions. The outer function is $g(x)$; the inner function is $h(x)$
.

When we can see $f(x)$ as the composite of two functions as in this case, the chain rule is:

If

$$f(x) = g(\ h(x)\ ),$$

then

$$f'(x) = g'(\ h(x)\ )\ h'(x).$$

In words we might say the derivative of a composite function is the derivative of the outer function times the derivative of the inner function. If the function is the composite $f(x) = g(h(x))$ given above, then we have:

$$
\begin{aligned}
f(x) &= (x^6 + 8)^4 \\
f'(x) &= 4(x^6 + 8)^3 \cdot (6x^5 + 0) \\
&= 4(x^5 + 8)^3 \cdot 6x^5 \\
&= 24x^5(x^5 + 8)^3
\end{aligned}
$$

The process is not difficult if you can see the function as inner and outer functions. Here are some further examples of the use of the chain rule:

Function	Derivative
$f(x) = (x + x^2)^4$	$f'(x) = 4(x + x^2)^3\ (1 + 2x)$
$f(x) = (x^3 - x^7)^2$	$f'(x) = 2(x^3 - x^7)^1(3x^2 - 7x^6)$
$f(x) = (x^3)^2$	$f'(x) = 2(x^3)^1(3x^2) = 6x^5$
$f(x) = (3x^4 - 2x^3 + x)^5$	$f'(x) = 5(3x^4 - 2x^3 + x)^4(12x^3 - 6x^2 + 1)$

Finding maximum and minimum points of functions

Finding maximum or minimum points of a function can then be placed in a simple set of rules. The rules given are appropriate for even more complex functions than we have discussed here. You would simply have to learn more methods of differentiation.

To find maximum and minimum values of a function:

1. Take the derivative of the function.

2. Set the derivative equal to zero and solve the resulting equation for the value of the domain of the function. Any point found may be a maximum, a minimum, or neither.

3. Evaluate the derivative to the left and right of any points found in step 2. If the derivative is positive to the left of the point and negative to the right, the function has a maximum at that point. If the derivative is negative to the left and positive to the right, the function has a minimum at that point. If the derivative is positive on both sides or negative on both sides, the point is neither a maximum or a minimum.

The reason for the last step are apparent after a moments thought. If the derivative is positive to the left and negative to the right, that means the function was increasing on the left of the point and decreasing on the right of the point. It was going up then down, it must have reached a maximum in between. Similarly, negative then positive means a minimum was reached. It is left as an exercise to decide what both positive or both negative values mean.

Let us see an example of this set of rules. Consider the function $f(x) = x^2 - 3x$. The following is the procedure.

Example of Finding a Maximum/Minimum Value

	$f(x) = x^2 - 3x.$	The original function.
Step 1:	Take the derivative.	
	$f'(x) = 2x - 3$	
Step 2:	Set the first derivative equal to zero and solve for x.	
	$2x - 3 = 0$	
	$2x = 3$	
	$x = 3/2$	
Step 3:	Evaluate the first derivative to the left and right of the point found in step 2.	
	$f'(1) = 2(1) - 3 = -1$	
	$f'(2) = 2(2) - 3 = 1$	

The derivative is negative to the left and positive to the right, the function has a minimum at 3/2.

These same steps will work to find maximums and minimums for many functions. Textbooks in calculus and mathematical handbooks give derivatives for many common forms of functions.

Partial derivatives

Functions can be defined on two variables. For example, the area of a rectangle is a function of the length and width of the rectangle,

$$A = l \cdot w .$$

This can be written as a function of two variables as

$$f(l,w) = l \cdot w .$$

Similarly, other functions are defined on two or more variables, some examples are

$$f(x,y) = (x + 2y)^3, \text{ and}$$

$$f(x,y) = (x - y) + 2xy \; .$$

In order to find minima and maxima of such functions we frequently take derivatives of the functions with respect to one variable at a time. We may take the derivative of the function with respect to one variable while considering the other variable to be a constant. This is called taking a partial derivative. The notation for a partial derivative is

$$\frac{\partial f}{\partial x},$$

which is read as "the partial derivative of f with respect to x".

When the partial derivative is taken twice with respect to two different variables, we have two derivatives both of which we set equal to zero. We then find the simultaneous solution of the two equations. At the simultaneous solution, we have the maximum value of the function with respect to both variables.

This approach can be used to find the least squares regression line. The regression line is an equation of the form

$$y' = a + bx \; .$$

Here, x is an observed value of the independent variable, y is an observed value of the dependent variable, y' is the estimated value of y, a is the intercept of the regression line and b is the slope of the regression line. What we would like to do is find estimators of a and b which will give us a best line. Best is usually defined as the least squares line, that is the line which will minimize the sum of the squared differences between y and y'. The sum of the squared differences is

$$\Sigma(y - y')^2,$$

which can also be written as

$$\Sigma(y - a - bx)^2$$

by substituting for y'.

We are used to thinking of this as a function of x and y, but we can see this last version can also be viewed as a function of a and b. We have then

$$f(a,b) = \Sigma(y - a - bx)^2.$$

Remember that the sigma notation simply indicates summation and that the derivative of a sum is the sum of the derivatives. We can thus take derivatives inside the summation and they will add appropriately:

$$\frac{\partial f}{\partial a} = \Sigma\ 2(y - a - bx)^1\ (-1)$$

$$\frac{\partial f}{\partial b} = \Sigma\ 2(y - a - bx)^1\ (-x).$$

If we set these two derivatives equal to zero and solve the resulting two equations simultaneously, we find the familiar formulas for simple linear regression[3]:

$$b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$a = \bar{y} - b\,\bar{x}$$

[3] The complete solution of these equations is in Appendix A. It would help understanding of this material considerably to go over the solution in detail.

Exercise Set 2

1. Plot the following functions and draw tangent lines to the curves at the maximum and minimum points.

   a) $f(x) = 2x^2 + 4$

   b) $f(x) = x^3 - 2x^2 + 4$

2. Find the derivatives of the following functions.

   a) $f(x) = 2.7182818285$

   b) $f(x) = 3x$

   c) $f(x) = 3x + 4$

   d) $f(x) = ex^2 + ex + e$

   e) $f(a) = (y - xa)^2$

   f) $f(b) = (y - b)^2$

   g) $f(x) = (y + x^2 - 4x^5)^4$

   h) $f(x) = 8 + ax^4 + (a - x^2)^2$

   i) $f(x) = 3.5x + (a + bx^4)^3$

3. Find the extreme values of the following functions.

   a) $f(x) = 5 - x$

   b) $f(x) = 3x - 2$

   c) $f(x) = x^2$

   d) $f(x) = x^3$

   e) $f(x) = x^2 - x + 2$

   f) $f(x) = 5x^2 - 2x + 1$

   g) $f(x) = x^3 - 4x$

   h) $f(x) = (x^2 - x)^2$

   i) $f(x) = (x^3 - x^2)^2$

   j) $f(x) = (x^2 - 1)^2$

3. What has been found when the first derivative is positive on both sides of a critical point? negative on both sides?

Integral Calculus

In integral calculus we find the area under a curve defined by a function. The process is referred to as integrating the function or we are said to find the integral. The sign for integration is

$$\int .$$

We usually write

$$\int f(x)\ dx$$

to indicate an indefinite integral. An indefinite integral is one in which we are only interested in finding the formula for the area, not an actual area. The dx which appears in the notation is called the differential. One of its purposes is to indicate the variable of integration, x in this case. More about the dx later, for the moment it is enough to simply notice that it is there.

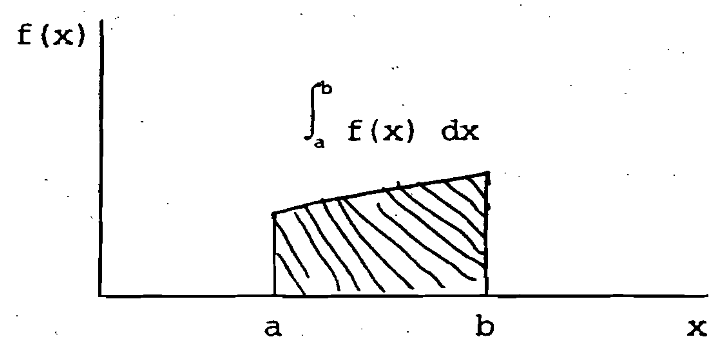
A definite integral is one in which we actually want the area. When we want the area, we indicate the bounds of the area we want to find. This is indicated by writing the definite integral with the limits for the integral:

$$\int_a^b f(x)\ dx \quad .$$

Here, the a is referred to as the lower limit of integration and the b as the upper limit of integration. This is probably most easily understood by looking at a drawing. Figure 8 shows the area of a function between the limits a and b; the area is marked with diagonal lines.

Figure 8

Integral of f(x) from the Lower Limit, a, to the Upper limit, b.



To this point we have seen the notation for integration and the concept of integration as area under a curve. How we actually find an integral is a bit surprising[4]. Integration turns out to be the reverse operation of differentiation. In a thorough discussion we would start by talking about the antiderivative. That is why differentiation is studied first. What we have to do is be able to remember what the process for differentiation was and reverse it. Lets look at what we do when there is a constant.

Integral of a constant

Consider a constant function,

$$f(x) = k.$$

The integral of a function in this form is

$$\int k \, dx = kx + c.$$

By the power rule, the derivative of a first power is

$$f'(x) = k \quad \text{when} \quad f(x) = kx.$$

Whatever constant appears as the coefficient of a first power of x will be left as the derivative when the power of x is reduced by one to zero. So, when we are integrating and we reverse the process, we simply put the variable of integration behind the constant.

This explains how where the kx comes from, where does the c come from? Remember that the derivative of a constant is 0. When we see a function such as $f(x) = k$, the 0 which could have been there from a differentiation is not written down. We have to remember that it is there and add an unknown constant, c. In many cases, c will turn out to be zero or to "vanish" from the solution.

Here are some examples:

---

4 Well, I was surprised when I saw it. If you have a normal amount of curiosity, you will wonder how this can all happen. How does this interesting and tidy relationship occur between these apparently different concepts? Once again, the calculus series in the mathematics department will answer many of these questions.

Function	Integral
$f(x) = 1$	$\int 1\,dx = 1x = x + c$
$f(x) = 4$	$\int 4\,dx = 4x + c$
$f(x) = 26$	$\int 26\,dx = 26x + c$
$f(x) = 4.3$	$\int 4.3\,dx = 4.3x + c$

Integral of a Power

You will recall that the derivative of a power is taken by multiplying the function by the power and then reducing the power by one. The converse is to increment the power by one and then to divide the function by the new power. Equivalently, we can multiply the function by 1 over the new power.
The formula for the general case is:

$$\int ax^n dx = \frac{ax^{n+1}}{n+1} + c$$

Some examples:

Function	Integral
$f(x) = x$	$\int x\,dx = 1/2\ x^2 + c$
$f(x) = x^4$	$\int x^4\,dx = 1/5\ x^5 + c$
$f(x) = x^6$	$\int x^6\,dx = 1/7\ x^7 + c$
$f(x) = 3x^4$	$\int 3x^4\,dx = 1/5 \cdot 3x^5 = 3/5\ x^5 + c$
$f(x) = 2.1x^6$	$\int 2.1x^6\,dx = 1/7 \cdot 2.1x^7 = .3x^7 + c$

Integral of a sum

As you probably would suspect from knowing the rules for differentiation, the integral of a sum is the sum of the integrals.

$$\int f(x) + g(x)\ dx = \int f(x)\ dx + \int g(x)\ dx .$$

We can use this with the previous two formulas to find integrals of some more complex functions:

Function	Integral
$f(x) = x + 1$	$\int x + 1\ dx = \int x\ dx + \int 1\ dx$
	$= 1/2\ x^2 + x + c$
$f(x) = 5x^4 + x$	$\int 5x^4 + 1\ dx = \int 5x^4\ dx + \int x\ dx$
	$= 1/5 \cdot 5x^5 + 1/2\ x^2 + c$
	$= x^5 + 1/2\ x^2$

Finding a Definite Integral

To find the definite integral, we evaluate the indefinite integral at the upper and lower limits of integration and subtract the resulting value for the lower limit from the upper limit. This is easiest seen by example:

$$\int_0^5 x\ dx = 1/2\ x^2 + c \Big|_0^5 = [1/2(5)^2 + c] - [1/2(0)^2 + c]$$

$$= 12.5 + c - [\quad 0 \quad + c]$$

$$= 12.5$$

At the first step in this problem, the vertical line with the upper and lower limits of integration at the right of the line indicates that the result is evaluated as the difference between the expression evaluated at the upper and lower limits. In other words, you evaluate the expression at the upper limit and then subtract the expression evaluated at the lower limits.

Exercise Set 3

1. Plot the following function on graph paper and find the area between 1 and 3 by counting squares and parts of squares.

$$f(x) = 3x$$

2. Find the following integral and compare your answer to that of problem 1.

$$\int_1^3 3x \, dx$$

3. Find the following definite integrals,

a) $\int_0^4 x \, dx$

b) $\int_9^{15} 1/2x \, dx$

c) $\int_5^{10} 1/2x^2 \, dx$

d) $\int_0^8 2x^3 - x^2 + x \, dx$

e) $\int_0^{10} x^4 - 2x^3 \, dx$

f) $\int_0^{20} 5x^4 - 2x \, dx$

g) $\int_0^5 1/5 \, x \, dx$

Statistics

Random Variables

A random variable is a number which is associated with the outcome of an experiment. Random variable can be discrete or continuous. It is assumed here that you have studied random variables before and that what is really necessary is to point out a few random variables, showing some which are discrete and some which are continuous. Accordingly, examples of both of these are given after their definitions.

Discrete random variables are random variables which can take on a countable number of points. For example:

- The number of heads which a coin shows when tossed one time is a discrete random variable which can take on two values, 0 and 1.
- The number of heads which a coin can show when tossed 10 times is a discrete random variable which can take on values 0, 1, 2, ..., 10.
- The number of men on a subcommittee of size three which is selected at random from a committee with fourteen members, including 5 women, is a discrete random variable which can take on values 0, 1, 2, and 3.
- The item score when a student answers an item on a test is a discrete random variable taking on values 0 or 1.
- The score which occurs when a student takes a test of 50 items is a discrete random variable which can take on values 0, 1, 2, ..., 50.

This last example may give you some pause since we generally treat test scores as continuous random variables. We do, but they aren't. We usually assume that the continuous distribution is a good approximation to the distribution of the test score. This is typically a good assumption.

Continuous random variables are random variables which are continuous functions as discussed previously. The definition will serve for our purposes. Some examples:

- If an experiment involving heights of first grade students, the height of first grade students is a continuous random variable which takes on values on the interval between 1 foot and 7 feet. (I am being conservative with the interval because I do not know the lower and upper limits of first grade students height.)

- In a study of smoking, the amount of nicotine a subject consumes in a day is a continuous random variable which takes on values between 0 and xx milligrams.

- The life of a light bulb is a continuous random variable taking on values between 0 and an upper limit which depends on the type bulb.

- The time a student spends doing homework is a continuous random variable taking on values between 0 and some upper limit which we will not attempt to specify here.

<u>Probability functions, cumulative probability functions, density functions, cumulative density functions and joint distributions functions.</u>

One of the first distinctions to make in these terms is that the terms probability function and cumulative probability function are used to refer to a discrete random variable while probability density function and cumulative density function refer to continuous random variables[5]. The term probability function refers to the function which relates the probability of a discrete random variable to the random variable itself. The cumulative probability function relates the sum of the probabilities of a random variable where the sum is taken from the lowest possible value of the random variable up to the given point. One of the most commonly studied discrete probability functions is the binomial. The binomial probability distribution function has some interest to us because it describes the number of success when the answers to a number of items are guessed.

A binomial experiment consists of n independent trials, each resulting in success or failure. The probability of success on each trial is p. We are interested in the total number of success in the n trials. The number of successes is the binomial random

---

[5] Okay, *most* authors use the terms in this manner. Some authors are not too careful, but you can usually determine what is happening from context.

variable. The formula for the probability function of a binomial random variable is

$$b(x;n,p) = C^n_x \; p^x \; q^{n-x}.$$

Here $q = 1 - p$ and C^n_x is the number of combinations of n things taken x at a time.

If we consider a ten item multiple choice test with 4 choices per item and we are interested in the number of correct answers when a student guesses, this is a binomial random variable with P = .25 and n = 10. Table 2 gives the probability distribution function and the cumulative probability distribution function for this binomial random variable. Table 2 is constructed by evaluating the formula for the binomial random variable for values of x ranging from 0 to 10. Those values are placed in the second column and the second column is cumulated to make the third column.

Other common discrete probability distributions are the hypergeometric and the poison distributions.

Table 2

Probability Distribution and
Cumulative Probability Distribution function
for a Binomial Random Variable with 10 trials and P = .25

x	P(X = x)	Cumulative Probability
0	.0563	.0563
1	.1877	.2440
2	.2816	.5256
3	.2503	.7759
4	.1460	.9219
5	.0548	.9803
6	.0162	.9965
7	.0029	.9996
8	.0004	1.0000
9	.0000	1.0000
10	.0000	1.0000

Continuous random variables have probability density functions, PDF's. These PDF's are analogous to the probability distribution functions of discrete random variables. It's the term analogous that causes the problem here. The problem is further compounded if we can not go far into a calculus based explanation. Let's see how far we can go.

A density function does not give the probability of the random variable but, as the name implies, it gives the density. The density is the height of the curve at any given value of the random variable. The density can be used to obtain the probability that the random variable falls within given limits. This is done by finding the area under the density curve between the given limits. Thus the density is closely related to probability, and we can say that it is analogous to the discrete probability distribution function, but it is not the same.

In introductory statistics courses, areas under the normal curve are found. Of course the hard part has already been done and the results tabled. The beginning student simply reads the tables. The process uses the cumulative density function. The area up to some limit or between some limits is used. In other words, what you have been doing since that first brush with statistics when you found normal curve areas is using a cumulative density function. Table 3 gives the values of the probability density function and the cumulative density function of the standard normal curve (μ =

0, $\sigma^2 = 1$) for selected values of the standard normal deviate z.

Table 3

Probability density function and
Cumulative Probability Density function
for a Normal Random Variable with $\mu = 0$, $\sigma^2 = 1$

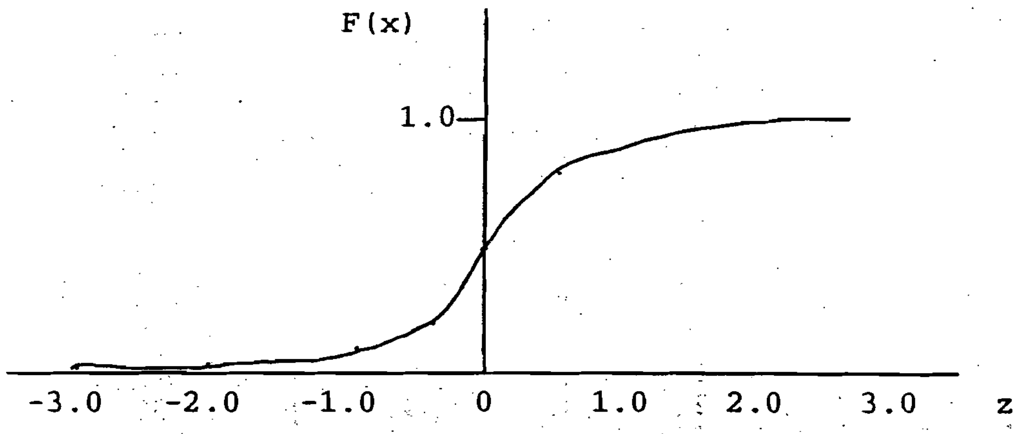
x	f(x)	Cumulative density function
-3.5	.0009	.0002
-3.0	.0044	.0013
-2.5	.0175	.0062
-2.0	.0540	.0238
-1.5	.1295	.0668
-1.0	.2420	.1587
-.5	.3107	.3085
0.0	.3989	.5000
.5	.3521	.6915
1.0	.2420	.8413
1.5	.1295	.9332
2.0	.0540	.9772
2.5	.0175	.9938
3.0	.0044	.9987
3.5	.0009	.9998

In Figure 9, the cumulative density function values from Table 3 have been plotted. Notice that the curve approaches but never reaches the value of 1. The curve is said to be "asymptotic" to 1 and 1 is said to be an asymptote of the curve.

The curve in Figure 9 is one of the main reasons we have been going through this review of mathematics and statistics. The form of the curve is what is important to us. The particular form of the curve, its characteristic "stretched out S" shape is what is noteworthy. That shape, or sometimes a portion of that shape, is what is found for many item response curves. As item response theory was developed, workers in the field attempted to derive methods which would yield that curve. Unfortunately, integrating the normal curve is not a straightforward process. In fact, simple formulas for normal curve areas do not exist and normal curve integrals are only found by approximation. The approximations are not readily usable in item response theory. No one was able to derive any theory with the cumulative normal density function. It was, as mathematicians say, intractable[6]. So what to do? Use something else of course; and that, dear reader, is the topic of a later section.

Figure 9

Plot of the cumulative density function of the standard normal curve



The Logistic Cumulative Distribution Function

Since those who developed item response theory were trained in statistics, it is only to be expected that they would want to use the cumulative normal distribution to approximate a curve that looks that in Figure 9. Due to the mathematical difficulties of working with the cumulative normal distribution, this did not prove feasible. A second candidate for use proved more easily managed, the logistic cumulative distribution function:

$$f(x) = \frac{1}{1 + e^{-x}}.$$

This function has the same general shape as that in Figure 9 and it proved to be much easier to work with than the cumulative normal distribution. The function has become central in item response theory.

---

6 My thesaurus gives these synonyms for intractable: disobedient, incorrigible, rebellious, uncontrollable, unmanageable, unruly, froward, obdurate, stubborn, uncooperative, uncompromising, unsubmissive, unyielding. It's easy to see that they mean the thing isn't going to work easily.

Exercise Set 4

1. Verify three of the probabilities on Table 2 by calculation.

2. Find f(x) for 3 of the x's on table 3. Compare your results with the values in the table.

3. Verify the cumulative probability for x = 3 on table 3 by comparing to a table you already know well.

4. Evaluate the cumulative logistic distribution function on the interval from -3.5 to 3.5 at every unit and half unit value. Plot the results and compare to Figure 9.

5. Say a continuous probability function is defined by the density function

   $f(x) = 2x/5, \quad 0 <= x <= 5$

   a) Plot f(x) over the interval [0,5].

   b) Use the formula for the area of a triangle to find the area under f(x) between 0 and 1.

   c) Find the integrals and graph the following areas.

   i) $\int_0^1 f(x)\,dx$

   ii) $\int_0^{2.5} f(x)\,dx$

   iii) $\int_1^{2.5} f(x)\,dx$

   iv) $\int_{2.5}^{5} f(x)\,dx$

6. If you know calculus well enough (more than we have done here), use Simpson's rule to estimate the area under the normal curve between z = -3 and z = -2. How good is the estimate?

7. Again, if you have a good calculus background, find the section on the normal curve in a mathematical statistics book such as those by Hogg and Craig or Mood and Graybill and follow the proof that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1$$

Maximum Likelihood Estimators

To the beginning student in statistics, it seems logical to estimate the population mean with the sample mean; the population variance with the sample variance; and so on. To the mathematical statistician, these estimators cry out for a methodology which justifies their use and will yield estimators when the case is not so obvious. Mathematical statisticians look for a procedure which will yield an estimator. For example, in simple linear regression how does one obtain the estimates of the slope and intercept. These estimators are not at all obvious as the sample mean is an obvious estimator of the population mean.

The slope and intercept in simple regression are found by the least squares approach[7]. The least squares approach is a procedure which yields estimators which minimize the squared distance between the estimate and the thing estimated. Mathematical statistics uses many least squares estimators and there is a fairly standard manner of finding least squares estimators. When a new parameter is defined and it is desired to estimate that parameter, a statistician can follow the least squares procedure to see if an estimator can be derived.

Another common procedure for finding an estimator is the method of maximum likelihood. In this procedure, a likelihood function is defined and then the maximum value of the function is found for some given sample data. The values which maximize the likelihood function are referred to as maximum likelihood estimators.

Sometimes the least squares estimator and the maximum likelihood estimator are identical, as when it is desired to find an estimator for the mean. Sometimes these estimators differ, as when it is desired to find an estimator for the variance. Other criteria are also important in selecting an estimator. Is the estimator consistent? Does it have minimum variance? These and other criteria are among the methods used to select estimators.

Maximum likelihood estimators are frequently easy to find and make good examples because of the ease of the theory. First we must define a likelihood function. Say that we have a random sample of size n from some distribution of known form. The likelihood function is the product of the n terms that result when the distribution function is evaluated at each observation in the function and each of these is multiplied together.

---

[7] See Appendix A for the derivation of the least squares estimates of the slope and intercept of a simple linear regression line.

Let's see the motivation for this definition. Since the n observations are random, they are independent. Independence of two random variables means that the joint probability function of the two random variables is the product of the probability functions of the two random variables. Thus, due to the independence, the probability of the joint occurrence of 2 of these observations is given by this product of the distribution functions. The argument extends to the general case of n random variables.

A particulary easy maximum likelihood estimator to find is the maximum likelihood estimate of the mean in a normal distribution. Remember the equation of a normal curve is

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

If we have a sample of x1, x_2, ... x_n, then the likelihood function is the product of the normal distribution evaluated at each of these observations. If we label the likelihood function L, we have

$$L = \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{1}{2}(\frac{x_1-\mu}{\sigma})^2} \cdot \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{1}{2}(\frac{x_2-\mu}{\sigma})^2} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi\sigma^2)}} e^{-\frac{1}{2}(\frac{x_n-\mu}{\sigma})^2}$$

Notice that this is the product of n terms, one term for each of the n observations in the sample. Further notice that each term consists of the product of a fractional constant and a power of e. Since there are n constant fractions we can gather them together and use an exponent to simplify the expression:

$$L = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \; e^{-\frac{1}{2}(\frac{x_1-\mu}{\sigma})^2} \cdot e^{-\frac{1}{2}(\frac{x_2-\mu}{\sigma})^2} \cdot \ldots \cdot e^{-\frac{1}{2}(\frac{x_n-\mu}{\sigma})^2}$$

Next, we notice that the powers of e can be combined by adding the exponents so that we obtain:

$$L = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \; e^{-\frac{1}{2}(\frac{x_1-\mu}{\sigma})^2 - \frac{1}{2}(\frac{x_2-\mu}{\sigma})^2 - \ldots - \frac{1}{2}(\frac{x_n-\mu}{\sigma})^2}$$

Now, to find the maximum likelihood estimator of μ, we have to maximize this function with respect to μ. With our meager knowledge of calculus, this looks like a formidable task, until we learn the tricks (Yes, two genuine tricks to make life easier). We reason thus:

> First, the fractional constant can be dropped since the function will be maximized with or without the constant.
>
> Second, if L is to be maximized, then the logarithm of L will also be maximized. Why not take the logarithm of L and maximize that.

First, we drop the constant leaving

$$L = e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2 - \frac{1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2 - \cdots - \frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2}$$

Next we take the logarithm. Obviously, this will be easiest if we use natural logarithms.

$$\ln(L) = -\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2 - \frac{1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2 - \cdots - \frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2$$

We then have

$$f(\mu) = -\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2 - \frac{1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2 - \cdots - \frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2$$

Again the function will be maximized if we factor out the fraction -½ and delete it for a constant multiplier will not change the maximum of the function.

$$f(\mu) = \left(\frac{x_1-\mu}{\sigma}\right)^2 + \left(\frac{x_2-\mu}{\sigma}\right)^2 + \cdots + \left(\frac{x_n-\mu}{\sigma}\right)^2$$

or

$$f(\mu) = \left(\frac{x_1}{\sigma}-\frac{\mu}{\sigma}\right)^2 + \left(\frac{x_2}{\sigma}-\frac{\mu}{\sigma}\right)^2 + \cdots + \left(\frac{x_n}{\sigma}-\frac{\mu}{\sigma}\right)^2$$

The derivative with respect to μ is then easily found by the chain rule

$$f' = 2(\frac{x_1}{\sigma}-\frac{\mu}{\sigma})(-\frac{1}{\sigma}) + 2(\frac{x_2}{\sigma}-\frac{\mu}{\sigma})(-\frac{1}{\sigma}) + \ . \ . \ . + 2(\frac{x_n}{\sigma}-\frac{\mu}{\sigma})(-\frac{1}{\sigma})$$

$$-\frac{2}{\sigma^2}(x_1-\mu) + (x_2-\mu) + \ . \ . \ . + (x_n-\mu) = 0$$

We set the derivative equal to zero and factor out the constants:

Then we multiply both sides by $-\sigma^2/2$ to remove the constant

$$(x_1-\mu) + (x_2-\mu) + \ . \ . \ . + (x_n-\mu) = 0$$

Rearranging terms we have

$$x_1 + x_2 + \ . \ . \ . + x_n - \mu - \mu - \ . \ . \ . - \mu = 0$$

or

$$\Sigma x_i - n\mu = 0$$

$$n\mu = \Sigma x_i$$

$$\mu = \Sigma x_i/n$$

Which is the maximum likelihood estimator for μ.

In this case, we would obtain the same estimator if we sought the least squares estimator. The maximum likelihood estimator and the least squares estimator sometimes produce the same estimate and sometimes do not produce the same estimate.

Exercise Set 5

1. State the form of the probability distributions functions for each of the discrete random variables on page 26.

2. Describe the procedure for obtaining a maximum likelihood estimator.

3. Reread the section on finding the maximum likelihood estimator of the mean, then take the likelihood function in its final form, take the derivative, and find the estimator. In other words, find the maximum likelihood estimator beginning at the point

$$f(\mu) = +(x_1/\sigma-\mu/\sigma)^2+(x_2/\sigma-\mu/\sigma)^2+\ldots+(x_n/\sigma-\mu/\sigma)^2 .$$

IMPORTANT TERMS

You should watch for these terms in the text and be sure you understand them.

abscissa
calculus
concave down
concave up
concavity
continuous function
continuous random variable
cumulative density function
cumulative probability function
density function
definite integral
dependent variable
derivative
differentiate
differential calculus
discrete random variable
domain
extrema
e
function
indefinite integral
independent variable
integral
integral calculus
irrational number
least squares estimator
logarithm
maximum likelihood estimator
ordinate
probability function
random variable
range
rational number
real number
relative maximum
relative minimum
tangent line

Appendix A

Finding least squares estimators for simple linear regression.

The regression line is an equation of the form

$$y' = a + bx .$$

Here, x is an observed value of the independent variable, y is an observed value of the dependent variable, y' is the estimated value of y, a is the intercept of the regression line and b is the slope of the regression line. What we would like to do is find estimators of a and b which will give us a best line. Best is usually defined as the least squares line. The least squares line is the line which will minimize the sum of the squared differences between y and y'. The sum of the squared differences is

$$\Sigma(y - y')^2,$$

which can also be written as

$$\Sigma(y - a - bx)^2.$$

We want to minimize this sum of squares. This is where the name for the "least squares approach" comes from.

We are used to thinking of this quantity to be minimized,

$$\Sigma(y - a - bx)^2,$$

as a function of x and y, but we can see this last version can also be viewed as a function of a and b. We have then

$$f(a,b) = \Sigma(y - a - bx)^2.$$

Remember that the sigma notation simply indicates summation and that the derivative of a sum is the sum of the derivatives. We can thus take derivatives inside the summation and they will add appropriately:

$$\frac{\partial f}{\partial a} = \Sigma\ 2(y - a - bx)^1\ (-1)$$

$$\frac{\partial f}{\partial b} = \Sigma\ 2(y - a - bx)^1\ (-x).$$

If we set these two derivatives equal to zero and solve the resulting two equations simultaneously, we find:

$\Sigma\ 2(y - a - bx)^1\ (-1) = 0$	Partial with respect to a.
$\Sigma\ 2(y - a - bx)^1\ (-x) = 0$	Partial with respect to b.
$-2\Sigma\ (y - a - bx) = 0$	Move constants outside Σ.
$-2\Sigma\ (y - a - bx)\ (x) = 0$	Move constants outside Σ.
$\Sigma\ (y - a - bx) = 0$	Divide by -2.
$\Sigma\ (y - a - bx)\ (x) = 0$	Divide by -2.
$\Sigma y - \Sigma a - \Sigma bx = 0$	Distribute Σ.
$\Sigma\ (xy - ax - bx^2) = 0$	Multiply through by x.
$\Sigma y - na - \Sigma bx = 0$	Summation rule.
$\Sigma xy - \Sigma ax - \Sigma bx^2 = 0$	Distribute Σ.
Equation 1: $\Sigma y - na - b\Sigma x = 0$	Move constant outside Σ.
Equation 2: $\Sigma xy - a\Sigma x - b\Sigma x^2 = 0$	Move constant outside Σ.

We now solve the Equation 1 for a and substitute that result for a in Equation 2 and solve the resulting equation.

$\Sigma y - na - b\Sigma x = 0$	Equation 1.
$na = \Sigma y - b\Sigma x$	Move terms across equal sign.
$a = \dfrac{(\Sigma y - b\Sigma x)}{n}$	Divide by n.

Substituting this result for a into Equation 2 we find:

$\Sigma xy - a\Sigma x - b\Sigma x^2 = 0$	Equation 2.
$\Sigma xy - \dfrac{(\Sigma y - b\Sigma x)\Sigma x}{n} - b\Sigma x^2 = 0$	Substitution.
$\Sigma xy - \underline{\Sigma x\Sigma y - b\Sigma x\Sigma x} - b\Sigma x^2 = 0$	Remove parentheses.

n

$$\Sigma xy - \frac{\Sigma x \Sigma y}{n} - \frac{b(\Sigma x)^2}{n} - b\Sigma x^2 = 0$$ Separate fraction into two fractions

$$-\frac{b(\Sigma x)^2}{n} + b\Sigma x^2 = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$$ Move terms across equal sign.

$$b\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right) = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$$ Factor b out of two terms.

$$b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$ Divide both sides by $\Sigma x^2 - \frac{(\Sigma x)^2}{n}$

This is the solution for the slope, b. To find the solution for a we can take equation 1 and solve for a.

$$\Sigma y - na - b\Sigma x = 0$$ Equation 1

$$\frac{\Sigma y}{n} - a - \frac{b\Sigma x}{n} = 0$$ Divide by n.

$$a = \frac{\Sigma y}{n} - \frac{b\Sigma x}{n}$$ Move terms across equal sign.

Recognizing the formulae for $\bar{y}$ and $\bar{x}$, we can rewrite the last equation as:

$$a = \bar{y} - b\bar{x} .$$