Chapter 2

Review of Mathematics

In this chapter, we will briefly review some of the mathematical concepts used in this textbook. Knowing these concepts will make it much easier to understand the mathematical underpinnings of statistics, especially the formulas used in statistics as well as their derivations. A particularly important concept is that of a function. We will commonly encounter several types of functions in statistics, including probability densities or distributions, likelihood functions, the functions used in statistical models, and ones used to transform the observations before statistical analysis.

2.1   Exponents

This section provides a brief summary of useful rules concerning exponents that often appear in statistical functions. Let a and b be two real numbers (numbers of any kind between $-\infty$ and ∞) that form the base of the exponent. This includes the special numbers $e \approx 2.71828$ and $\pi \approx 3.14159$ that often occur in statistics. As exponents or powers, let m and n be any positive

integers $(1, 2, 3, ...)$. We then have

$$a^m a^n = a^{m+n} \tag{2.1}$$
$$(a^m)^n = a^{mn} \tag{2.2}$$
$$\frac{a^m}{a^n} = a^{m-n} \tag{2.3}$$
$$(a \times b)^n = a^n b^n \tag{2.4}$$
$$\left(\frac{a}{b}\right)^n = \frac{a^n}{b^n} \tag{2.5}$$

(Schmidt & Ayres 2003). For example, suppose that $a = 2$, $b = 3$, $m = 5$, and $n = 4$. We have

$$a^m a^n = a^{m+n} \tag{2.6}$$
$$2^5 2^4 = 2^{5+4} = 2^9 = 512 \tag{2.7}$$

$$(a^m)^n = a^{mn} \tag{2.8}$$
$$(2^5)^4 = 2^{5\times4} = 2^{20} = 1048576 \tag{2.9}$$

$$\frac{a^m}{a^n} = a^{m-n} \tag{2.10}$$
$$\frac{2^5}{2^4} = 2^{5-4} = 2^1 = 2 \tag{2.11}$$

$$(a \times b)^n = a^n b^n \tag{2.12}$$
$$(2 \times 3)^5 = 2^5 3^5 = 32 \times 243 = 7776 \tag{2.13}$$

$$\left(\frac{a}{b}\right)^n = \frac{a^n}{b^n} \tag{2.14}$$
$$\left(\frac{2}{3}\right)^5 = \frac{2^5}{3^5} = \frac{32}{243} = 0.132 \tag{2.15}$$

These rules also hold for m and n any real number provided a and b are positive. Some special cases of the above rules are also commonly encountered

in statistics. We have

$$a^0 = 1, a \neq 0 \tag{2.16}$$
$$0^0 = 0 \tag{2.17}$$
$$a^{1/2} = \sqrt{a} \tag{2.18}$$
$$a^{-m} = \frac{1}{a^m}, a \neq 0 \tag{2.19}$$

Now suppose that $a = 4$ and $m = 2$. We have

$$4^0 = 1 \tag{2.20}$$
$$4^{1/2} = \sqrt{4} = 2 \tag{2.21}$$
$$4^{-2} = \frac{1}{4^2} = \frac{1}{16} = 0.0625 \tag{2.22}$$

2.2 Inequalities

Statistical statements often involve the use of inequalities. For example, suppose that you are interested in the size distribution of a fish population. You might be interested in estimating the probability or proportion of fish that equal or exceed the legal catch size, say 12 inches. If y stands for fish size, then you would be interested in estimating the probability of fish for which $y \geq 12$ inches. You might also be interested in fish which lie within a certain range of size, say 6 to 12 inches. This could be written as $6 < y < 12$ inches using inequalities. The results of statistical tests are often reported using inequalities as well. You will commonly encounter statements of the form '$P < 0.05$' in scientific papers, which says that the probability P of a certain event occurring is less than 5%, or 1 chance in 20.

Inequalities can be manipulated much like equalities in algebra, with some exceptions. Let x and y stand for any two numbers, or more complex mathematical quantities. If $x < y$, then

$$x + b < y + b \tag{2.23}$$

where b is another number or quantity, and

$$ax < ay \tag{2.24}$$

where a is a **positive** number or other quantity. If a is **negative**, then

$$ax > ay. \tag{2.25}$$

Thus, multiplying an inequality by a negative number flips the direction of the inequality. For example, let $x = 5$, $y = 6$, and $a = -2$. We have $x < y$, but clearly $-2(5) = -10$ is greater than $-2(6) = -12$.

Another exception involves the inverse or reciprocal of an inequality. If $x < y$ and both are positive (or both negative), then

$$\frac{1}{x} > \frac{1}{y}. \tag{2.26}$$

Note the changed direction of the inequality. For example, if $x = 5$ and $y = 6$ so that $x < y$, the inequality is reversed because we have $1/5 > 1/6$. However, if $x < y$ and x is negative, then

$$\frac{1}{x} < \frac{1}{y}. \tag{2.27}$$

For example, if $x = -5$ and $y = 6$ then we have $1/-5 < 1/6$, or $-1/5 < 1/6$. These results can also be obtained through direct application of Eq. 2.24 and 2.25.

2.3   Functions

A variable is a symbol such as x or y chosen to represent a set of numbers, typically real numbers. A function is a relationship between x and y such that each value of x generates a single value of y (Schmidt & Ayres 2003). When such a relationship holds, it is customary to say that y is a function of x. An example of a function is the equation

$$y = 2x + 1 \tag{2.28}$$

This happens to be the equation of a line with a slope of 2 and an intercept of 1. In general, we can write a function using the notation

$$y = f(x) \tag{2.29}$$

where $f(x)$ stands for any possible function of x. In this context, x is often called the independent variable and y the dependent variable.

2.3.1   Functions in Statistics

One commonly used function in statistics is the equation for a line, namely

$$y = ax + b \tag{2.30}$$

where a is the slope and b is the intercept of the line. This function plays an important role in linear regression, a statistical procedure that fits a line to a series of points of the form (x, y) (see Chapter 17). Also common are quadratic functions of the form

$$y = ax^2 + bx + c \tag{2.31}$$

where a, b, and c are constants. Rather than a straight line, quadratic functions are shaped like a parabola.

Exponential and log functions are also commonly used in statistics. Examples of exponential functions are

$$y = 10^x \tag{2.32}$$

and

$$y = e^x, \tag{2.33}$$

where $e = 2.71828\ldots$, also written as

$$y = \exp(x). \tag{2.34}$$

Examples of log functions are the natural log and base 10 log, written as

$$y = \ln(x) \tag{2.35}$$

and

$$y = \log(x). \tag{2.36}$$

Confusingly, the natural log is sometimes written as $\log(x)$, while base 10 log is written as $\log_{10}(x)$. SAS uses this notation for log functions. The log functions are only defined for $x > 0$.

The exponential and log functions are inverses, meaning they reverse the action of each other. For example, we have

$$\exp(\ln(x)) = x \tag{2.37}$$

and

$$\ln(\exp(x)) = x. \tag{2.38}$$

For example, if you find $\ln(x)$ for some value of x, then apply the exp function to $\ln(x)$, you get the original value of x as the answer. Suppose that $x = 2$. We have $\ln(x) = \ln(2) = 0.693$, and then $\exp(\ln(2)) = \exp(0.693) = 2$. The same thing happens for the functions 10^x and $\log(x)$.

Another common function in statistics is the absolute value function, written as

$$y = |x|. \tag{2.39}$$

It is defined as follows. If x is positive or zero then $|x|$ is simply equal to x, while if x is negative then $|x| = -x$. For example, if $x = -2$ then $y = |-2| = -(-2) = 2$. A common use of the absolute value in statistics is to define a symmetric interval around zero. For example, the inequality $-3 < x < 3$ can also be written as $|x| < 3$.

The most commonly used distribution in statistics is the normal distribution, which can be written as a combination of several simpler functions:

$$y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.40}$$

Here μ and σ^2 are two parameters that govern the shape of the normal distribution, in particular its mean and variance (see Chapter 6).

2.3.2   Plotting functions using SAS - SAS demo

It can be difficult to discern the shape of a function without a graph. For example, the function describing the normal distribution gives you the famous bell-shaped curve, but this is not obvious from the equation. We will develop a SAS program that will plot any function, given its mathematical form, the values of any constants, and the range of x values for which a plot is needed. We will examine this plotting program in some detail, because it illustrates the structure of the programs used throughout this textbook.

SAS programs consist of a series of steps or instructions that enable you to input and manipulate data and then generate statistical results and graphs. Data are entered and manipulated using SAS `data` steps, while statistical results and graphs are generated using SAS procedures or `proc` steps. Note that SAS is not case-sensitive, so programs can be in either upper or lower case.

The first line of the program is a comment, used here to give the file name of the program. Any line of a SAS program beginning with an asterisk (*) is a comment, which are used to describe the program and its actions but are not executed by SAS.

```
* fplot.sas;
```

The next three lines consist of the instructions

```
title "Plot a function y = f(x)";
title2 "Linear function";
```

The two `title` lines add a main title and subtitle to the output. Note that each of the lines ends with a semicolon (;). This is absolutely critical in SAS programming, because it tells SAS where a particular statement or command ends. A misplaced or absent semicolon will typically cause errors when running the program.

The next part of the SAS code is a `data` step (SAS Institute Inc. 2016a). The idea here is to generate a data set with a sequence of x and $y = f(x)$ values that will later be plotted. The minimum and maximum values of x are set by specifying values for `xmin` and `xmax`, while the number of divisions is set by `xdiv` (the more divisions the finer the x scale and the smoother the graph). The program then calculates the step length between x values (`xlength`) using these quantities. The values of x and $y = f(x)$ are calculated in a programming loop using a `do` statement . Each pass through the loop calculates a new value of x, then finds $y = f(x)$ for that value of x. The results are then sent to a SAS data file using an `output` statement. You can set the name of the data file in the first line of the `data` step, which in this case is `fplot`. Note that six different functions are listed in this `data` step, but only one would be active (the line function) because the remainder are comments. This is a useful programming trick to deactivate sections of code.

```
data fplot;
    * Minimum and maximum values of x;
    xmin = -5;
    * Use for ln function, must have x > 0;
    *xmin = 0.001;
    xmax = 5;
    * Divisions between xmin and xmax (more = smoother graph);
    xdiv = 100;
    * Calculate step length;
    xlength = (xmax-xmin)/xdiv;
    * Find x and y = f(x) values for the plot;
    do i=0 to xdiv;
        x = xmin + i*xlength;
        * Insert f(x) formula here;
        * line function;
        y = 2*x + 1;
        * quadratic function;
        *y = -x**2 + 2*x + 5;
        * exponential function;
        *y = exp(x);
        * ln function;
        *y = log(x);
        * absolute value function;
        *y = abs(x);
        * normal distribution;
        *mu = 1;
        *sig2 = 1;
        *y = (1/sqrt(2*3.14159*sig2))*exp(-((x-mu)**2)/(2*sig2));
        * Output x and y to SAS data file;
        output;
    end;
run;
```

The resulting data are then printed using the SAS `print` procedure (SAS Institute Inc. 2016b), using the syntax below. The option `data=fplot` tells the `print` procedure to use this particular data file. If this option were omitted, the last data file created would automatically be used. The `run` statement tells SAS that the `proc print` command is complete and that it should get busy printing the data file.

```
* Print data;
proc print data=fplot;
run;
```

The `gplot` procedure is used to plot the function using the new data set (see below) (SAS Institute Inc. 2016c). The `plot` statement tells SAS which of your SAS variables are the x and y variables - the variable before the asterisk (`*`) is the y variable, after it the x variable (it is the position that is important, not the name of the variable). The `href = 0` and `vref = 0` options make SAS draw vertical and horizontal lines through the origin $(0, 0)$. The `symbol1` statement tells SAS to join the points with a line (`i=join`), draw no symbol for each data point (`v=none`), and make the line connecting the points red (`c=red`). The remainder of the options listed in the program are intended to make the graph more legible by increasing the thickness of the lines and size of the axes labels. If you are curious how they work, try experimenting with the numbers given in the options. The `quit` statement returns control to SAS after running the program.

```
* Plot y = f(x);
proc gplot data=fplot;
    plot y*x=1 / href=0 vref=0 whref=3 wvref=3 vaxis=axis1 haxis=axis1;
    symbol1 i=join v=none c=red width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```

See the full program listing below, a portion of the printed output, and graphs for the various functions included in the program. Both the program and complete output can be found on the website for this textbook.

```
――――――――――――――――――――――――――――― SAS program ―――――――――――――――――――――――――

* fplot.sas;
title "Plot a function y = f(x)";
title2 "Linear function";
data fplot;
    * Minimum and maximum values of x;
    xmin = -5;
    * Use for ln function, must have x > 0;
    *xmin = 0.001;
    xmax = 5;
    * Divisions between xmin and xmax (more = smoother graph);
    xdiv = 100;
    * Calculate step length;
    xlength = (xmax-xmin)/xdiv;
    * Find x and y = f(x) values for the plot;
    do i=0 to xdiv;
        x = xmin + i*xlength;
        * Insert f(x) formula here;
        * line function;
        y = 2*x + 1;
        * quadratic function;
        *y = -x**2 + 2*x + 5;
        * exponential function;
        *y = exp(x);
        * ln function;
        *y = log(x);
        * absolute value function;
        *y = abs(x);
        * normal distribution;
        *mu = 1;
        *sig2 = 1;
        *y = (1/sqrt(2*3.14159*sig2))*exp(-((x-mu)**2)/(2*sig2));
        * Output x and y to SAS data file;
        output;
    end;
run;
* Print data;
proc print data=fplot;
run;
* Plot y = f(x);
proc gplot data=fplot;
    plot y*x=1 / href=0 vref=0 whref=3 wvref=3 vaxis=axis1 haxis=axis1;
    symbol1 i=join v=none c=red width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
```

```
run;
quit;
```

Obs	xmin	xmax	xdiv	xlength	i	x	y
Plot a function y = f(x)							
Linear function							
1	-5	5	100	0.1	0	-5.0	-9.0
2	-5	5	100	0.1	1	-4.9	-8.8
3	-5	5	100	0.1	2	-4.8	-8.6
4	-5	5	100	0.1	3	-4.7	-8.4
5	-5	5	100	0.1	4	-4.6	-8.2
6	-5	5	100	0.1	5	-4.5	-8.0
7	-5	5	100	0.1	6	-4.4	-7.8
8	-5	5	100	0.1	7	-4.3	-7.6
9	-5	5	100	0.1	8	-4.2	-7.4
10	-5	5	100	0.1	9	-4.1	-7.2

etc.

Figure 2.1: `fplot.sas` - `proc print`
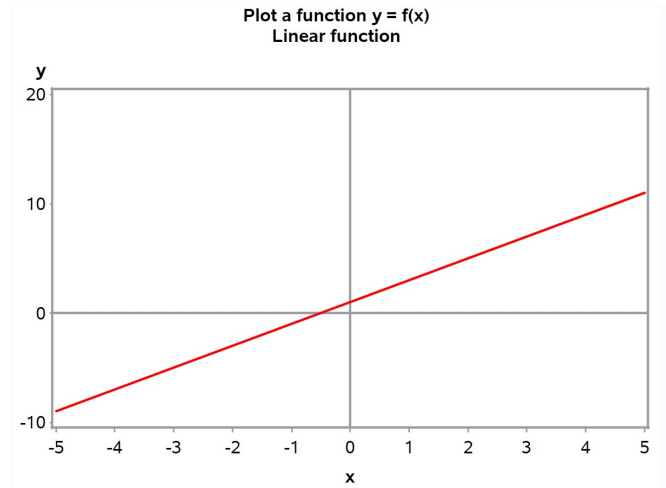
Figure 2.2: `fplot.sas - proc gplot`

Figure 2.3: `fplot.sas - proc gplot`

Plot a function y = f(x)
Exponential function

Figure 2.4: `fplot.sas - proc gplot`

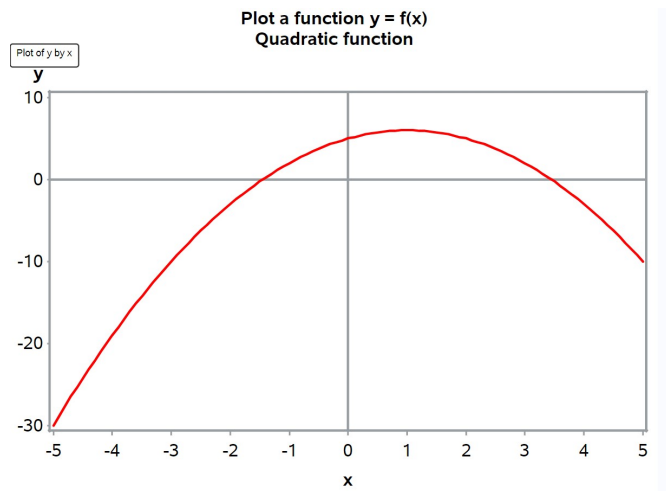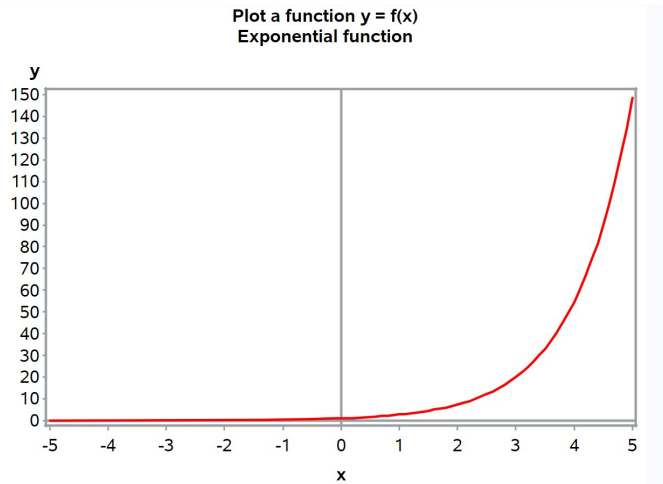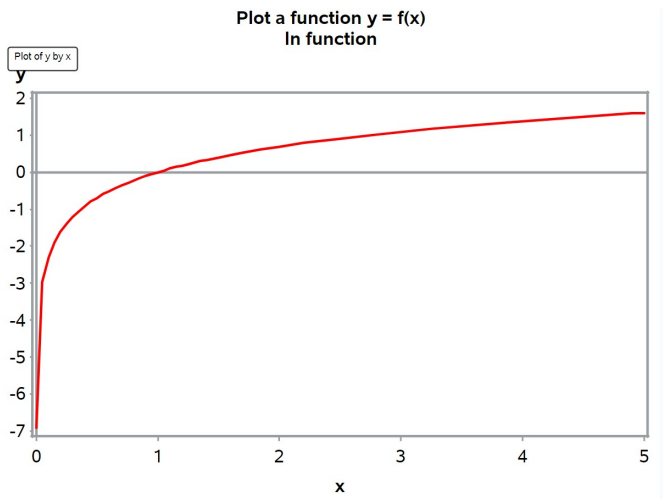**Plot a function y = f(x)**
ln function
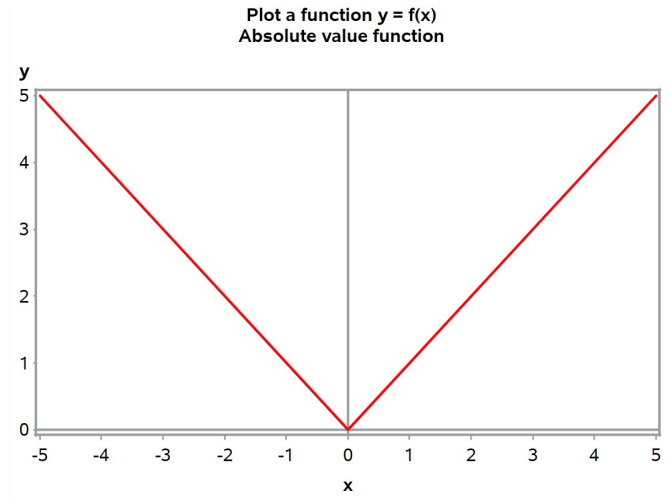
Figure 2.5: `fplot.sas - proc gplot`
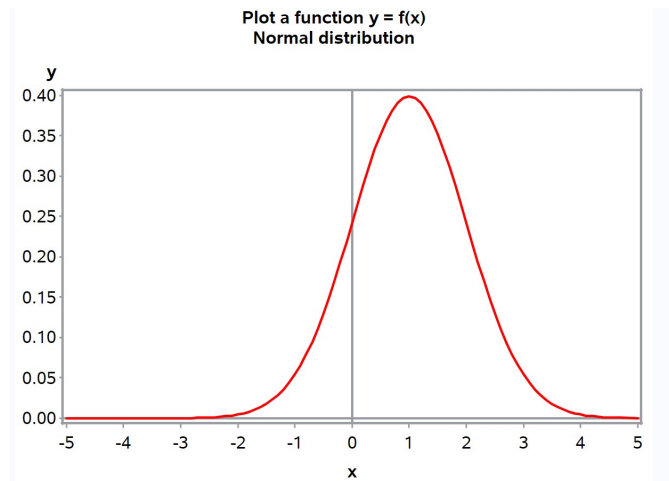
Figure 2.6: `fplot.sas - proc gplot`

Figure 2.7: `fplot.sas - proc gplot`

2.4 Solving linear equations

We next review how to solve a linear equation for x, a procedure that will be useful in later developments. A linear equation has the general form

$$ax + b = cx + d \tag{2.41}$$

where a, b, c, and d are constants that are possibly zero, while x is a variable. We want to find a value of x that makes this equation true, meaning the two sides of the equation are equal. To solve this problem, you perform the same operations on both sides of the equation until you have x alone on one side of the equation. The other side is then the answer to this problem (Schmidt & Ayres 2003). More generally, a-d and x could also be more complicated expressions that one manipulates to obtain an expression for x.

To illustrate this procedure, suppose we have the equation

$$5x - 4 = 3x - 3. \tag{2.42}$$

Subtracting $3x$ from both sides of the equation, we get

$$2x - 4 = -3. \tag{2.43}$$

We next add 4 to both sides to obtain

$$2x = 1. \tag{2.44}$$

Dividing both sides by 2 we obtain the solution

$$x = 1/2. \tag{2.45}$$

If you want to check if the solution is correct, you can always substitute it back into the original equation. We have

$$5(1/2) - 4 = 3(1/2) - 3 \tag{2.46}$$
$$2.5 - 4 = 1.5 - 3 \tag{2.47}$$
$$-1.5 = -1.5. \tag{2.48}$$

So $x = 1/2$ is in fact the correct solution.

2.5   Roots of equations

For a particular function $y = f(x)$, it is often useful to find the values of x for which $y = f(x) = 0$. Values of x for which this is true are called the roots of the equation $f(x) = 0$ (Schmidt & Ayres 2003). Graphically, the roots are the values of x where the function crosses the x-axis, i.e., the function is equal to zero. It is possible to find the roots for many functions algebraically, but not every function has roots, and for some functions they can only be found numerically using software and a computer.

Roots are easy to find for linear functions. Recall that a linear function takes the general form

$$y = a + bx \tag{2.49}$$

where a and b are constants. We want to find values of x for which

$$a + bx = 0 \tag{2.50}$$

We then use the rules for solving linear equations to find x. Subtracting a from both sides and dividing by b, we obtain

$$x = \frac{-a}{b} \tag{2.51}$$

Suppose that $a = 1$ and $b = 2$, so that our function is

$$y = 1 + 2x. \tag{2.52}$$

It follows that the root of this function is $x = -a/b = -1/2$. If we examine the graph generated earlier for this function, we see that the function indeed crosses the x-axis at $x = -1/2$.

We can also find the roots for quadratic functions using, logically enough, the quadratic formula. Recall that a quadratic function takes the general form

$$y = ax^2 + bx + c \tag{2.53}$$

We want to find values of x for which

$$ax^2 + bx + c = 0. \tag{2.54}$$

The quadratic formula says that the roots are given by the equation

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{2.55}$$

We previously plotted a quadratic function of the form

$$y = -x^2 + 2x + 5 \tag{2.56}$$

To find the roots, we need to solve the equation

$$-x^2 + 2x + 5 = 0 \tag{2.57}$$

Inspecting this equation, we see that $a = -1$, $b = 2$, and $c = 5$. Inserting these values in the quadratic formula, we obtain

$$x = \frac{-2 \pm \sqrt{2^2 - 4(-1)5}}{2(-1)} = \frac{-2 \pm \sqrt{24}}{-2} \tag{2.58}$$

$$= \frac{-2 \pm 4.90}{-2} = \frac{-6.90}{-2}, \frac{2.90}{-2} = 3.45, -1.45 \tag{2.59}$$

The roots of this quadratic equation are therefore equal to 3.45, 1.45. This result agrees with the graph drawn earlier.

2.6 Calculus

We will make only limited use of calculus in this course, but it is useful to review the concepts of derivatives and integrals. Derivatives are often used in estimating the parameters of statistical models through a method called maximum likelihood (Chapter 8). Integrals are used to generate the probabilities associated with confidence intervals, statistical tests, and other procedures. For example, the statistical tables given in Chapter 23 were all generated using integrals.

2.6.1 Derivatives

A derivative of a function $y = f(x)$ is defined to be the slope of the function at a particular value of x. Recall that the slope is defined as the change in y divided by the change in x. The mathematical definition of a derivative is given by the equation

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \tag{2.60}$$
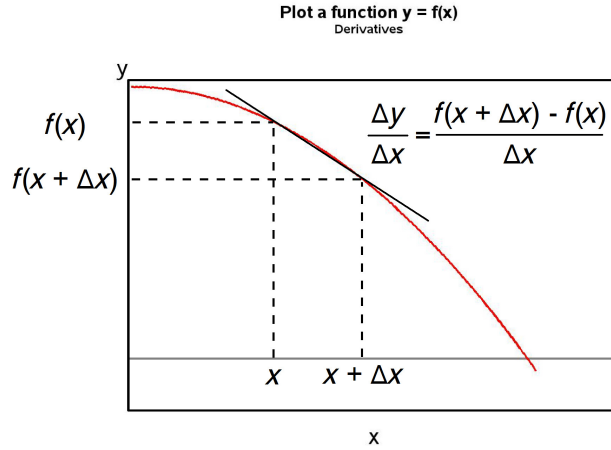
Figure 2.8: Definition of a derivative

where Δx is the change in x, while Δy is the change in y, defined as $f(x + \Delta x) - f(x)$ (Schmidt & Ayres 2003). This equation says that the derivative is given by the limit, as Δx goes to zero, of the slope $\Delta y / \Delta x$. See also Fig. 2.8. The derivative of a function may be written as $\frac{dy}{dx}$ or $f'(x)$.

Now suppose we have a linear function like

$$y = ax + b. \tag{2.61}$$

The derivative of this function is simply a, the slope of the line. It is equal to a regardless of the value of x, because a line has the same slope everywhere. We would write this as $\frac{dy}{dx} = a$ or $f'(x) = a$.

Assume now that we have a quadratic function. There is a formula for the derivative of a power of x that is often useful. If $y = f(x) = kx^n$, where k and n are any constants, then

$$\frac{dy}{dx} = knx^{n-1}. \tag{2.62}$$

We can use this formula to find the derivative of a quadratic function of the form

$$y = ax^2 + bx + c. \tag{2.63}$$

We have

$$\frac{dy}{dx} = a(2)x^{2-1} + b(1)x^{1-1} + 0 = 2ax + b. \tag{2.64}$$

To obtain this result, we also made use of the fact that the derivative of a constant (c in this case) is always zero (because it is unchanging), and that the derivative of a sum of functions is the sum of the derivatives.

One important application of the derivative in statistics is to find the maximum or minimum of a function. In particular, the derivative of a function is equal to zero at the maximum or minimum. This follows because a function that has a maximum must eventually stop rising and begin to fall, and at that point the slope is equal to zero. The same reasoning applies to a minimum.

To find the maximum or minimum for our general quadratic function, we set $dy/dx = 0$ and solve for x. We have

$$\frac{dy}{dx} = 2ax + b = 0. \tag{2.65}$$

Solving this linear equation for x, we find that the maximum or minimum will occur at $x = \frac{-b}{2a}$.

2.6.2   Function plot with derivative - SAS demo

We will plot a quadratic function and its derivative to observe the relationship between the two. Suppose that we have the following quadratic function:

$$y = -x^2 + 2x + 5. \tag{2.66}$$

The derivative of this function is

$$\frac{dy}{dx} = -2x^{2-1} + 2(1)x^{1-1} + 0 = -2x + 2. \tag{2.67}$$

We can find the minimum or maximum of this function by setting the derivative equal to zero and solving for x. We have

$$-2x + 2 = 0 \tag{2.68}$$

for which the solution is $x = 1$.

We will now plot both y and dy/dx using a revised version of our plotting program. This program calculates both y and dy/dx within the `do` loop,

then plots both sets of points on the same graph using the overlay option in `proc gplot`. See SAS program and output below.

Note that the derivative of this quadratic function is a straight line with a slope of -2 and an intercept of 2. It equals zero at the point where it intercepts the x-axis, which also corresponds to the maximum of the quadratic function. Our calculation above shows this occurs at $x = 1$.

———————————————— SAS program ————————————————

```
* fplot_deriv.sas;
title "Plot a function and its derivative";
title2 "Quadratic function";
data fplot2;
    * Minimum and maximum values of x;
    xmin = -5;
    xmax = 5;
    * Divisions between xmin and xmax (more = smoother graph);
    xdiv = 100;
    * Calculate step length;
    xlength = (xmax-xmin)/xdiv;
    * Find x, y = f(x), and dy/dx values for the plot;
    do i=0 to xdiv;
        x = xmin + i*xlength;
        * quadratic function;
        y = -x**2 + 2*x + 5;
        * derivative of this function;
        dydx = -2*x + 2;
        * Output x, y, and dydx to SAS data file;
        output;
    end;
run;
* Print data;
proc print data=fplot2;
run;
* Plot y = f(x) and dydx;
proc gplot data=fplot2;
    plot y*x=1 dydx*x=2 / href=0 vref=0 overlay whref=3 wvref=3 vaxis=axis1
    haxis=axis1;
    symbol1 i=join v=none c=red width=3;
    symbol2 i=join v=none c=blue width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```

Plot a function and its derivative
Quadratic function

Obs	xmin	xmax	xdiv	xlength	i	x	y	dydx
1	-5	5	100	0.1	0	-5.0	-30.00	12.0
2	-5	5	100	0.1	1	-4.9	-28.81	11.8
3	-5	5	100	0.1	2	-4.8	-27.64	11.6
4	-5	5	100	0.1	3	-4.7	-26.49	11.4
5	-5	5	100	0.1	4	-4.6	-25.36	11.2
6	-5	5	100	0.1	5	-4.5	-24.25	11.0
7	-5	5	100	0.1	6	-4.4	-23.16	10.8
8	-5	5	100	0.1	7	-4.3	-22.09	10.6
9	-5	5	100	0.1	8	-4.2	-21.04	10.4
10	-5	5	100	0.1	9	-4.1	-20.01	10.2

etc.

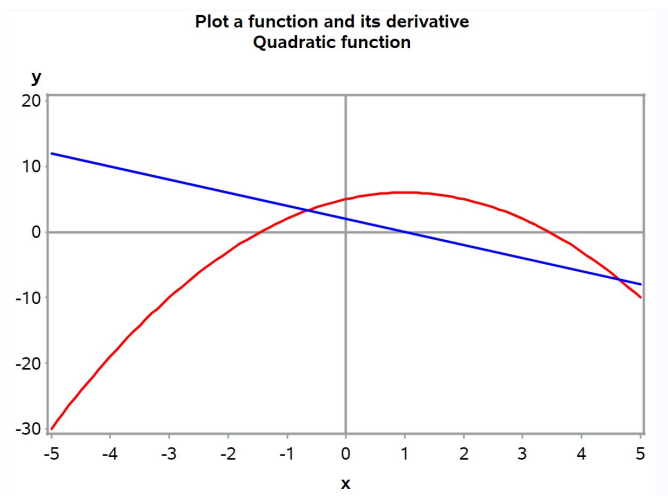Figure 2.9: `fplot_deriv.sas` - `proc print`

Figure 2.10: `fplot_deriv.sas` - `proc gplot`

2.6.3   Integrals

Statistics makes heavy use of integrals in working with the normal and other statistical distributions, although statistical tables or software typically do the work for the end user. For example, tables of the normal distribution provide probabilities for certain intervals - these probabilities are actually areas under the bell-shaped curve and are calculated by integration.

One kind of integral often encountered in statistics is a called a definite integral. It is basically the area A under a function $f(x)$ over some range of x values, say $a < x < b$. It is written mathematically as the equation

$$A = \int_a^b f(x)dx. \tag{2.69}$$

Here the symbol $\int$ is the integral sign, with the range of x values $(a < x < b)$ shown as sub- and superscripts of the integral sign.

To make things more concrete, we will illustrate definite integrals using the normal distribution function. Consider this function for $\mu = 5$ and $\sigma^2 = 1$, and the area A under it from $x = 5$ to $x = 6$ (Fig. 2.11). If we were modeling the behavior of some biological variable (say body mass of a small animal) using this distribution, the area A would be the probability that an animal falls within this range of x values. It would be expressed in mathematical terms as the integral

$$A = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_5^6 \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-5)^2}{2}} dx. \tag{2.70}$$

How is the area A actually calculated through integration? We can approximate this area by dividing it into strips of width $\Delta x = 0.25$ and height $f(x)$ given by the normal distribution function (Fig. 2.12). Adding the areas of these strip, we obtain $A \approx 0.099 + 0.093 + 0.080 + 0.068 = 0.340$. If we increased the number of strips while simultaneously decreasing the width of the strips Δx, we would get an even more accurate approximation to A. The integral is defined as the limit of this process, as the number of strips approaches infinity and their width $\Delta x \to 0$ (Schmidt & Ayres 2003). The exact value of the area obtained through this process is $A = 0.341$.
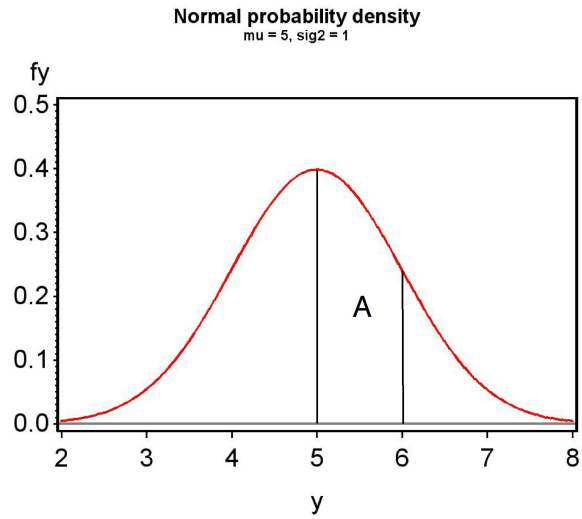
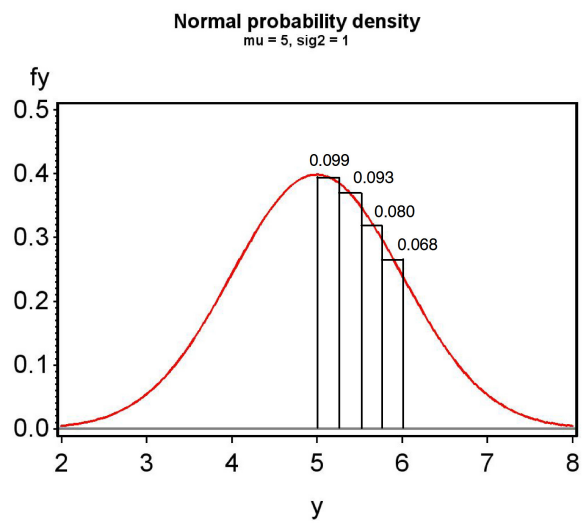Figure 2.11: Plot of the normal distribution for $\mu = 5$ and $\sigma^2 = 1$.

Figure 2.12: Plot of the normal distribution for $\mu = 5$ and $\sigma^2 = 1$.

2.7   References

SAS Institute Inc. (2016a) *SAS 9.4 Language Reference: Concepts, Sixth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2016b) *Base SAS 9.4 Procedures Guide, Sixth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2016c) *SAS/GRAPH 9.4: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

Schmidt, P. A. & Ayres, F. Jr. (2003) *Schaums Outline of Theory and Problems of College Mathematics, 3rd Edition.* The McGraw-Hill Companies, Inc., New York, NY.

2.8 Problems

1. Suppose that you have the quadratic function $y = x^2 - 2x - 8$. Find the roots of this function, then determine the value of x that minimizes it. Plot the function using SAS, attaching your program, output, and graph.

2. Consider the quadratic function $y = -2x^2 + 5x + 5$. Find the roots of this function, then determine the value of x that maximizes it. Plot the function and its derivative dy/dx using SAS, attaching your program, output, and graph.

3. Plot the function $y = 0.5\lambda^3 x^2 \exp(-\lambda x)$ for $\lambda = 2$ and $0 \leq x \leq 5$ using SAS. Attach your program, output, and graph. This function is a special case of the gamma distribution, a probability distribution often used to model continuous data (see Chapter 6).