The Logistic Function

R.E. Deakin
Bonbeach VIC, 3196, Australia
email: randm.deakin@gmail.com
October 2018

Introduction

The *logistic function* or *logistic curve* is a common S-shape curve (sigmoid curve) with equation:

$$P\left(t\right) = \frac{K}{1 + e^{-\left(rt + C\right)}} \tag{1}$$

where $P\left(t\right)$ is the population P at time t, K is the carrying capacity and the curve's maximum value, $e = 2.718281828\ldots$ is a mathematical constant and the base of the natural logarithms, r is the growth parameter (steepness of the curve), $C = \ln\left(\dfrac{P_0}{K - P_0}\right)$ is a constant and P_0 is the population at $t = 0$.

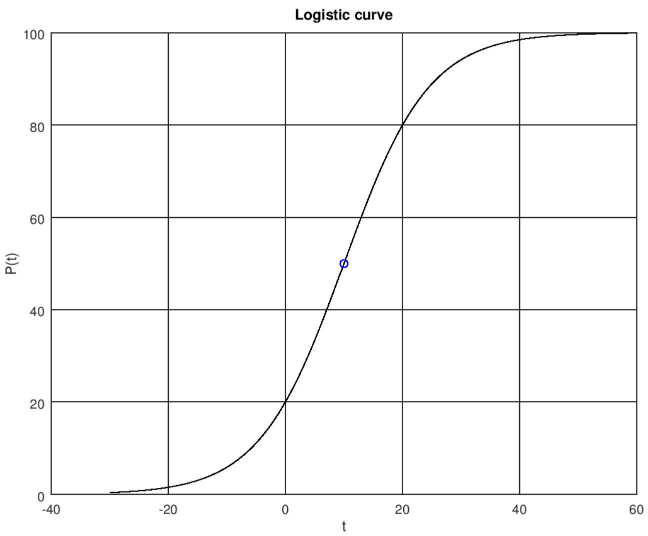

Figure 1. Logistic curve: $P\left(t\right) = \dfrac{K}{1 + e^{-\left(rt + C\right)}}$, $K = 100$, $P_0 = 20$, $r = 0.138629436$.

For real values of t in the range $-\infty < t < +\infty$, the curve has two asymptotes, $P\left(t\right) = K$ as $t \to +\infty$ and $P\left(t\right) = 0$ as $t \to -\infty$ and is symmetric about the curve's midpoint at $t = t_0 = -\dfrac{C}{r}$ where $P\left(t_0\right) = \frac{1}{2}K$.

There are various alternative expressions for the logistic curve and the derivation of (1) – outlined below – is a modern interpretation of the derivation by Pierre Verhulst in the 19th century who named the curve *la courbe logistique* [the logistic curve].

The properties of the logistic curve are derived and a general equation developed with a special case, the sigmoid curve. This is followed by the derivation of the *logistic distribution.*

Logistic regression is discussed and the method of *least squares* is employed to give a solution for the parameters of a logistic curve that is a best fit of the outcomes of binary dependent variables. Two examples of this technique are shown.

Finally, the connection between a sport rating system (Elo 1978) and the logistic curve is shown.

References for further reading and two appendices are included.

A Brief History of the Logistic Function

The development of the logistic function is due to Pierre François Verhulst (1804–1849) and his work on population growth in the 19th century. In an 1838 paper titled *Notice sur loi que la population suit dans son accroissement* [A Note on the Law of Population Growth], Verhulst proposed a differential equation that modelled bounded population growth and solved this equation to obtain the population function. He then compared actual populations with his modelled values for France (1817-1831), Belgium (1815-1833), Essex, England (1811-1831) and Russia (1796-1827). In a subsequent 1844 paper titled *Recherches mathématiques sur la loi d'accroissement de la population* [Mathematical research on the Law of Population Growth] he named the solution to the equation he had proposed in his 1838 paper *la courbe logistique* [the logistic curve].

Verhulst's original works on population modelling were criticised by others (and even Verhulst in an 1847 publication) and his work was largely ignored, but in the early 20th century the logistic function was 'rediscovered' and Verhulst acknowledged as its inventor. Since then the logistic function (and logistic growth) is a simple and effective model of population growth that is applied in many diverse fields of scientific study (Cramer 2002, O'Connor & Robertson 2014, Bacaër 2011).

Derivation of the Logistic Function

[The notation of Bacaër (2011) is used in this derivation.]

Verhulst proposed the following (somewhat arbitrary) differential equation for the population $P(t)$ at time t

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right) \tag{2}$$

where $P(t)$ is the population P at time t, K is the carrying capacity and r is the growth parameter.

The differential equation (2) for the population P is solved by integration where

$$\int \frac{dP}{P\left(1 - \frac{P}{K}\right)} = \int r\,dt \tag{3}$$

In order to evaluate the left-hand-side we write

$$\frac{1}{P\left(1 - \frac{P}{K}\right)} = \frac{K}{KP - P^2} = \frac{K}{P(K - P)}$$

and decomposing into partial fractions gives

$$\frac{K}{P(K-P)} = \frac{A}{P} + \frac{B}{K-P} = \frac{A(K-P) + BP}{P(K-P)}$$

from which the relation $A(K-P) + BP = K$ is obtained which must hold for all values of K and P. Choosing particular values for P yields A and B as follows: (i) when $P = 0$, $AK = K$ and $A = 1$; (ii) when $P = K$, $BK = K$ and $B = 1$. Using this result gives

$$\frac{1}{P\left(1 - \frac{P}{K}\right)} = \frac{K}{P(K-P)} = \frac{1}{P} + \frac{1}{K-P}$$

and (3) becomes

$$\int \frac{1}{P}\,dP + \int \frac{1}{K-P}\,dP = \int r\,dt \tag{4}$$

Using the substitution $u = K - P$ with $du = -dP$ in the second integral gives

$$\int \frac{1}{P}\,dP - \int \frac{1}{u}\,du = \int r\,dt$$

and using the integral results $\int \frac{1}{x}\,dx = \ln x + C$ and $\int a\,dx = ax + C$ where the C's are constants of integration gives

$$\ln P - \ln\left(K - P\right) = rt + C \tag{5}$$

where ln denotes the natural logarithm and $\ln x \equiv \log_e x$ and $e = 2.718281828\ldots$, and the constants of integration have been combined and added to the right-hand-side.

Re-writing (5) as

$$\ln\left(K - P\right) - \ln P = -\left(rt + C\right)$$

and using the laws of logarithms: $\log_a \frac{M}{N} = \log_a M - \log_a N$ gives

$$\ln \frac{K - P}{P} = -\left(rt + C\right)$$

Now, raising both sides to the base e and noting that $e^{\ln x} = x$ gives

$$\frac{K - P}{P} = e^{-\left(rt + C\right)} \tag{6}$$

that can be re-arranged as

$$P = P\left(t\right) = \frac{K}{1 + e^{-\left(rt + C\right)}} \tag{7}$$

An equation for the constant C can be found when $t = 0$ in which case $P_0 = P\left(0\right) = \frac{K}{1 + e^{-C}}$ and a re-arrangement gives

$$e^{-C} = \frac{K - P_0}{P_0} \quad \text{and} \quad e^{C} = \frac{P_0}{K - P_0} \tag{8}$$

Taking natural logarithms of both sides of the second member of (8) and noting $\ln e^x = x$ gives

$$C = \ln\left(\frac{P_0}{K - P_0}\right) \tag{9}$$

Properties of the Logistic Curve

Asymptotes

The curve described by (7) is a symmetric S shape (see Figure 1) with two asymptotes, the upper one being the line $P\left(t\right) = K$ as $t \to +\infty$ and the lower one being the line $P\left(t\right) = 0$ as $t \to -\infty$.

Midpoint

The midpoint of the curve is when $P\left(t\right) = \frac{1}{2}K$ and this occurs when the exponent $\left(rt + C\right)$ in (7) is equal to zero and

$$t = t_0 = -\frac{C}{r} \tag{10}$$

Thus the midpoint of the logistic curve is at $\left(t_0, \frac{1}{2}K\right)$

Symmetry

The logistic curve is symmetric about the midpoint.

This can be confirmed by writing the denominator of (7) as $1 + e^{-\left(rt+C\right)} = 1 + e^{-r\left(t+C/r\right)} = 1 + e^{-r\left(t-t_0\right)}$ giving

$$P = P\left(t\right) = \frac{K}{1 + e^{-r\left(t-t_0\right)}} \tag{11}$$

Now using (11) with $d = t - t_0$ a distance along the t-axis, the sum of the logistic function $P\left(d\right)$ and its reflection about the vertical axis through t_0, $P\left(-d\right)$ is

$$\begin{aligned}\frac{K}{1+e^{-rd}} + \frac{K}{1+e^{-\left(-rd\right)}} &= \frac{K\left(1+e^{rd}\right) + K\left(1+e^{-rd}\right)}{\left(1+e^{-rd}\right)\left(1+e^{rd}\right)} \\ &= \frac{2K + K\left(e^{rd} + e^{-rd}\right)}{2 + e^{rd} + e^{-rd} + e^{\left(-rd+rd\right)}} \\ &= \frac{K\left(2 + e^{rd} + e^{-rd}\right)}{2 + e^{rd} + e^{-rd}} \\ &= K\end{aligned}$$

Thus the logistic function is symmetric about the midpoint $\left(t_0, \frac{1}{2}K\right)$

Also, due to symmetry we may write $\dfrac{K}{1+e^{-r\left(t-t_0\right)}} = K - \dfrac{K}{1+e^{r\left(t-t_0\right)}}$ and (11) becomes

$$P = P\left(t\right) = K - \frac{K}{1 + e^{r\left(t-t_0\right)}} \tag{12}$$

Inflexion point

The midpoint is also the inflexion point of the logistic curve where the second derivative $\dfrac{d^2P}{dt^2} = 0$. This can be proved by the following

First, to simplify the analysis write the denominator of (7) as $1 + e^{-\left(rt+C\right)} = 1 + e^{-rt}e^{-C}$ and using the first member of (8) we let

$$A = e^{-C} = \frac{K - P_0}{P_0} \tag{13}$$

and (7) becomes

$$P = P\left(t\right) = \frac{K}{1 + Ae^{-rt}} \tag{14}$$

Differentiating (14) with respect to t gives

$$\frac{dP}{dt} = -K\left(1 + Ae^{-rt}\right)^{-2}\left(-Are^{-rt}\right)$$

and differentiating again gives

$$\frac{d^2P}{dt^2} = 2K\left(1+Ae^{-rt}\right)^{-3}\left(-Are^{-rt}\right)^2 - K\left(1+Ae^{-rt}\right)^{-2}\left(Ar^2e^{-rt}\right)$$

Second, solving $\frac{d^2P}{dt^2} = 0$ gives

$$\begin{aligned}
&2K\left(1+Ae^{-rt}\right)^{-3}\left(-Are^{-rt}\right)^2 - K\left(1+Ae^{-rt}\right)^{-2}\left(Ar^2e^{-rt}\right) = 0\\
&2\left(1+Ae^{-rt}\right)^{-1}\left(Are^{-rt}\right)^2 - rAre^{-rt} = 0\\
&2\left(1+Ae^{-rt}\right)^{-1}\left(Are^{-rt}\right) - r = 0\\
&2\left(1+Ae^{-rt}\right)^{-1}\left(Ae^{-rt}\right) = 1\\
&2Ae^{-rt} = 1 + Ae^{-rt}\\
&Ae^{-rt} = 1\\
&e^{-rt} = \frac{1}{A} = A^{-1}
\end{aligned}$$

Taking logarithms of both sides and noting that $\log_a M^p = p\log_a M$ gives

$$t = \frac{\ln A}{r} \tag{15}$$

Substituting (15) into (14) gives

$$P\left(\frac{\ln A}{r}\right) = \frac{K}{1+Ae^{-r\left(\frac{\ln A}{r}\right)}} = \frac{K}{1+Ae^{-\ln A}} = \frac{K}{1+Ae^{\ln A^{-1}}} = \frac{K}{1+AA^{-1}} = \tfrac{1}{2}K$$

Finally, using (9), (10) and (13) we obtain the relations

$$\ln A = -C \quad \text{and} \quad t_0 = \frac{\ln A}{r} \tag{16}$$

So, $P\left(\frac{\ln A}{r}\right) = P\left(t_0\right) = \tfrac{1}{2}K$ thus proving that the inflexion point is the midpoint $\left(t_0, \tfrac{1}{2}K\right)$.

Verhulst's equation for the logistic function

In Verhulst's 1838 paper he states the differential equation

$$\frac{dp}{dt} = mp - \varphi\left(p\right) \tag{17}$$

that links the rate of change of population p with respect to time t with mp and a function of p, $\varphi\left(p\right)$ where m is a constant. He then supposes that $\varphi\left(p\right) = np^2$ where n is another constant and then finds for the integral of (17) that

$$t = \frac{1}{m}\left[\ln p - \ln\left(m - np\right)\right] + \text{constant}$$

On resolving (17) he gives the equation for the population p as

$$p = \frac{mp'e^{mt}}{np'e^{mt} + m - np'} \tag{18}$$

where p' is the population at $t = 0$. He then states that as $t \to \infty$ the value of p corresponds with $P = \frac{m}{n}$ that he calls *la limite supérieure de la population* [the upper limit of the population].

The correspondence between variables in Verhulst's 1838 paper and this paper are

Verhulst	This paper	
p	P	population
t	t	time
m	r	growth parameter
n	no equivalent	
p'	P_0	population at $t = 0$
$P = \frac{m}{n}$	K	upper limit of population

Verhulst's differential equation (17) can be written as $\frac{dp}{dt} = mp - np^2 = mp\left(1 - \frac{n}{m}p\right)$ that is equivalent to $\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right)$ that is our equation (2) and Verhulst's logistic function (18) can be written as $p = \dfrac{p'e^{mt}}{1 + \frac{n}{m}p'\left(e^{mt} - 1\right)}$ that is equivalent to (Bacaër 2011, eq. 6.2)

$$P = \frac{P_0 e^{rt}}{1 + \frac{P_0}{K}\left(e^{rt} - 1\right)} = \frac{KP_0 e^{rt}}{K + P_0\left(e^{rt} - 1\right)} \tag{19}$$

[This equation can be obtained from (14) with a bit of algebra]

Verhulst's method for estimating the parameters *r* and *K*

In Verhulst's 1844 paper he set out a method of estimating the parameters r and K from the population $P(t)$ in three different but equally spaced years; P_0 at $t = 0$, P_1 at $t = t_1$ and P_2 at $t = t_2 = 2t_1$. Verhulst's method is as follows.

Using (6) the relationship $e^{(rt+C)} = \frac{P}{K - P}$ can be obtained and taking logarithms of both sides gives $rt + C = \ln\left(\frac{P}{K - P}\right)$ and using this equations at times $t = 0,\ t = t_1$ and $t = t_2 = 2t_1$ gives

$$\begin{aligned} C &= \ln\left(\frac{P_0}{K - P_0}\right) \\ rt_1 + C &= \ln\left(\frac{P_1}{K - P_1}\right) \\ 2rt_1 + C &= \ln\left(\frac{P_2}{K - P_2}\right) \end{aligned}$$

and from these, the following can be obtained

$$e^{C} = \frac{P_0}{K - P_0} \qquad (\mathrm{i})$$
$$e^{rt_1}e^{C} = \frac{P_1}{K - P_1} \qquad (\mathrm{ii})$$
$$e^{2rt_1}e^{C} = \frac{P_2}{K - P_2} \qquad (\mathrm{iii})$$

Dividing (ii) by (i) and (iii) by (ii) gives two equations where e^{C} has been eliminated and the left-hand-sides are identical

$$e^{rt_1} = \frac{P_1\left(K - P_0\right)}{P_0\left(K - P_1\right)} \qquad (\mathrm{iv})$$
$$e^{rt_1} = \frac{P_2\left(K - P_1\right)}{P_1\left(K - P_2\right)} \qquad (\mathrm{v})$$

Equating (iv) and (v) gives $\frac{P_1\left(K - P_0\right)}{P_0\left(K - P_1\right)} = \frac{P_2\left(K - P_1\right)}{P_1\left(K - P_2\right)}$ and cross-multiplying and gathering terms gives

$$\left(P_1^2 - P_0P_2\right)K^2 - \left(P_0P_1^2 + P_2P_1^2 - 2P_0P_1P_2\right)K = 0$$

Dividing both sides by K and rearranging gives (Bacaër 2011, p.38)

$$K = \frac{P_1\left(P_0P_1 + P_2P_1 - 2P_0P_2\right)}{P_1^2 - P_0P_2} \tag{20}$$

noting the conditions $P_1^2 > P_0P_2$ and $P_0P_1 + P_1P_2 > 2P_0P_2$ for finite and positive K.

Once K is known, r can be obtained from (iv) as

$$r = \frac{1}{t_1}\ln\left(\frac{P_1\left(K - P_0\right)}{P_0\left(K - P_1\right)}\right) = \frac{1}{t_1}\ln\left(\frac{1/P_0 - 1/K}{1/P_1 - 1/K}\right) \tag{21}$$

Logistic Function – Various forms

Various forms of the logistic function have been developed so far in this paper – see (7), (11), (12), (14) and (19)

$$P\left(t\right) = \frac{K}{1 + e^{-\left(rt + C\right)}} = \frac{K}{1 + e^{-r\left(t - t_0\right)}} = K - \frac{K}{1 + e^{r\left(t - t_0\right)}} = \frac{K}{1 + Ae^{-rt}} = \frac{P_0e^{rt}}{1 + \frac{P_0}{K}\left(e^{rt} - 1\right)} \tag{22}$$

where $P_0 = P\left(0\right)$, $C = \ln\left(\frac{P_0}{K - P_0}\right)$, $A = \frac{K - P_0}{P_0}$, $t_0 = -\frac{C}{r} = \frac{\ln A}{r}$,

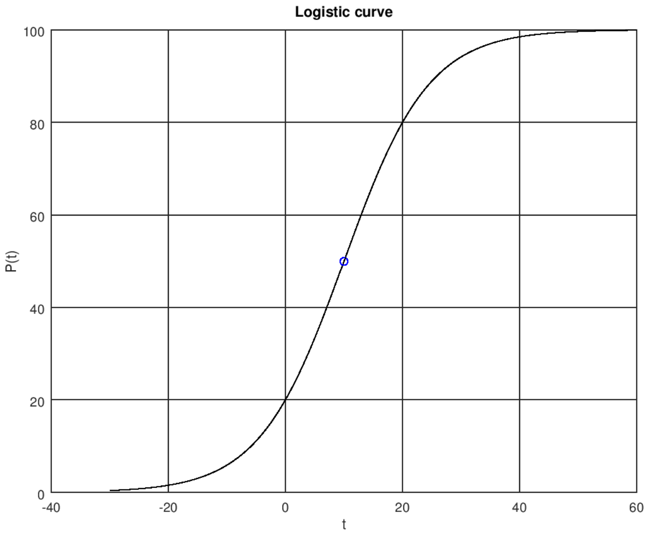

Figure 2. Logistic curve: $P\left(t\right) = \dfrac{K}{1 + e^{-r\left(t - t_0\right)}}$, $K = 100$, $t_0 = 10$, $r = 0.138629436$

Logistic Function – A General Form

Considering (12), a *general form* for the logistic function can be given as

$$y = \frac{A_1 - A_2}{1 + e^{k\left(x - x_0\right)}} + A_2 \tag{23}$$

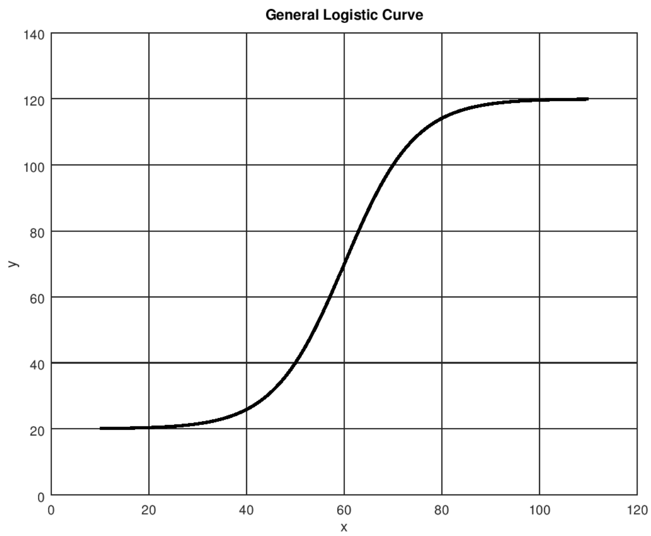

Figure 3. Logistic curve: $y = \dfrac{A_1 - A_2}{1 + e^{k\left(x - x_0\right)}} + A_2$, $A_1 = 20$, $A_2 = 120$, $k = 0.138629436$, $x_0 = 60$

$y = A_1$ is the lower asymptote of the curve given by (23) when $x \to -\infty$ and $y = A_2$ is the upper asymptote of the curve when $x \to +\infty$. The midpoint of the curve is $\left[x_0, \frac{1}{2}\left(A_1 + A_2\right)\right]$ when $x = x_0$ then $e^{k\left(x - x_0\right)} = e^0 = 1$ and $y = \frac{1}{2}\left(A_1 + A_2\right)$

Sigmoid Function

The sigmoid function is a special case of the logistic function. The sigmoid curve is a symmetric S-shape with a midpoint at $\left(0,\frac{1}{2}\right)$ and asymptotes $y=0$ as $x \to -\infty$ and $y=1$ as $x \to +\infty$.

$$y = \frac{1}{1+e^{-x}} = \frac{e^{x}}{e^{x}+1} \tag{24}$$

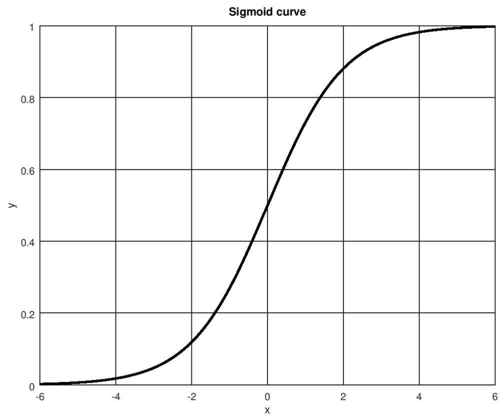

Figure 4. Sigmoid curve: $y = \dfrac{1}{1+e^{-x}}$

The derivative of the sigmoid function, denoted by $y' = \dfrac{dy}{dx}$ is

$$y' = \frac{d}{dx}\left(1+e^{-x}\right)^{-1} = -1\left(\left(1+e^{-x}\right)^{-2}\right)\left(-e^{-x}\right) = \frac{e^{-x}}{\left(1+e^{-x}\right)^{2}} \tag{25}$$

and from (24) $e^{-x} = \dfrac{1-y}{y}$ and $1+e^{-x} = \dfrac{1}{y}$ giving

$$y' = y\left(1-y\right) \tag{26}$$

The 2nd derivative of the sigmoid function, denoted by $y'' = \dfrac{d^2y}{dx^2} = \dfrac{d}{dx}\left(\dfrac{dy}{dx}\right)$ is

$$\begin{aligned} y'' &= \frac{d}{dx}\left(e^{-x}\left(1+e^{-x}\right)^{-2}\right) \\ &= e^{-x}\left(-2\left(1+e^{-x}\right)^{-3}\left(-e^{-x}\right)\right) + \left(1+e^{-x}\right)^{-2}\left(-e^{-x}\right) \\ &= \frac{e^{-x}}{\left(1+e^{-x}\right)^{2}}\left(\frac{2e^{-x}}{1+e^{-x}}-1\right) \\ &= y'\left(1-2y\right) \end{aligned} \tag{27}$$

And as before, the inflexion point of the sigmoid curve will be at the point where $y''=0$ and using (27) $y''=0$ when $y=\frac{1}{2}$ which occurs at the midpoint $\left(0,\frac{1}{2}\right)$

The Logistic Distribution

This section assumes some knowledge of statistical concepts and a brief outline of some of these is given in Appendix A.

Suppose X is a continuous random variable and $X = x$ denotes x as a possible real value of X.

The *probability density function* $f_X\left(x\right)$ has the following properties

$$1. \qquad f_X\left(x\right) \geq 0 \text{ for any value of } x \tag{28}$$

$$2. \qquad \int_{-\infty}^{+\infty} f_X\left(x\right) dx = 1 \tag{29}$$

The probability that a random variable X lies between any two values $x = a$ and $x = b$ is the area under the density curve between those two values and is found by methods of integral calculus

$$P\left(a < X < b\right) = \int_a^b f_X\left(x\right) dx \tag{30}$$

The *cumulative distribution function* $F_X\left(x\right)$ has the following properties

$$1. \qquad F_X\left(x\right) = P\left(X \leq x\right) = \int_{-\infty}^{x} f_X\left(x\right) dx \tag{31}$$

$$2. \qquad \frac{d}{dx} F_X\left(x\right) = f_X\left(x\right) \tag{32}$$

In many scientific analyses of experimental data it is assumed the data are members of a probability distribution with a density function having a smooth bell-shaped curve with tails that approach the asymptote $f_X\left(x\right) = 0$ as $x \to \pm\infty$ and a cumulative distribution function that is a symmetric S-shape with asymptotes $F_X\left(x\right) = 0$ and $F_X\left(x\right) = 1$ as $x \to -\infty$ and $x \to +\infty$ respectively.

As an example, Appendix A shows the probability density curve and the cumulative distribution curve of the familiar Normal distribution with probability density function $f_X\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ where the infinite population has mean μ, variance σ^2 and standard deviation $\sigma = +\sqrt{\sigma^2}$.

Now suppose that our experimental data has a *Logistic distribution* and the cumulative distribution function is the logistic function with location parameter a and shape parameter b [see (23) with $A_1 = 0$, $A_2 = 1$, $x_0 = a$ and $k = 1/b$]

$$F_X\left(x\right) = \frac{1}{1 + e^{-\left(\frac{x-a}{b}\right)}} \tag{33}$$

The probability density function $f_X\left(x\right)$ is given by (32) as

$$f_X\left(x\right) = \frac{d}{dx}\left[\frac{1}{1 + e^{-\left(\frac{x-a}{b}\right)}}\right] = \frac{d}{dx}\left[\left(1 + e^{-\left(\frac{x-a}{b}\right)}\right)^{-1}\right]$$

Let $u=-\left(\dfrac{x-a}{b}\right)$ and with $\dfrac{du}{dx}=-\dfrac{1}{b}$ we write $\dfrac{d}{dx}\left[\left(1+e^{-\left(\frac{x-a}{b}\right)}\right)^{-1}\right]=\dfrac{d}{du}\left[\left(1+e^{u}\right)^{-1}\right]\dfrac{du}{dx}$ and using the rule $\dfrac{d}{dx}e^{x}=e^{x}$ the probability density function of the Logistic distribution is

$$f_X\left(x\right)=\frac{e^{-\left(\frac{x-a}{b}\right)}}{b\left(1+e^{-\left(\frac{x-a}{b}\right)}\right)^2} \tag{34}$$

Mean and Variance of the Logistic Distribution

The mean μ_X and variance σ_X^2 are special mathematical expectations (see Appendix A) and

$$\mu_X=E\left\{X\right\}=\int_{-\infty}^{+\infty}x\,f_X\left(x\right)dx \tag{35}$$

$$\sigma_X^2=E\left\{\left(X-\mu_X\right)^2\right\}=\int_{-\infty}^{+\infty}\left(x-\mu_x\right)^2 f_X\left(x\right)dx \tag{36}$$

Using the rules for expectations a more useful expression for the variance can be developed as

$$\begin{aligned}\sigma_X^2&=E\left\{\left(X-\mu_X\right)^2\right\}=E\left\{X^2-2X\mu_X+\left(\mu_X\right)^2\right\}\\&=E\left\{X^2\right\}-2\mu_X E\left\{X\right\}+\left(\mu_X\right)^2=E\left\{X^2\right\}-2\mu_X\mu_X+\left(\mu_X\right)^2\\&=E\left\{X^2\right\}-\left(\mu_X\right)^2=E\left\{X^2\right\}-\left(E\left\{X\right\}\right)^2\end{aligned} \tag{37}$$

The following derivations are due to Max Hunter (2018), my mentor in all things mathematical especially the lovely integral solutions that follow[1].

The mean

Using (34) and (35) the mean can be written as

$$\mu_X=\int_{-\infty}^{+\infty}x\,f_X\left(x\right)dx=\int_{-\infty}^{+\infty}\frac{x\,e^{-\left(\frac{x-a}{b}\right)}}{b\left(1+e^{-\left(\frac{x-a}{b}\right)}\right)^2}dx \tag{38}$$

With the substitution $t=\dfrac{x-a}{b}$ then $x=tb+a$ and $dx=b\,dt$, and with $t=\infty$ when $x=\infty$ then (38) becomes

[1] Max Hunter is a retired mathematician from RMIT University, Melbourne Australia. In an earlier version of this document I had resorted to the use of a divergent series in solving integrals for the mean and variance. Max, on reading this, sent me a note suggesting I had set back mathematics and statistics about 300 years! But very kindly attached several elegant solutions that avoided the use of said series.

$$\mu_X = \int_{-\infty}^{+\infty} \frac{\left(tb + a\right)e^{-t}}{\left(1 + e^{-t}\right)^2} dt = b\int_{-\infty}^{+\infty} \frac{te^{-t}}{\left(1 + e^{-t}\right)^2} dt + a\int_{-\infty}^{+\infty} \frac{e^{-t}}{\left(1 + e^{-t}\right)^2} dt \tag{39}$$

Three results are useful in evaluating (39).

First

$$\frac{e^x}{\left(1 + e^x\right)^2} = \frac{e^x}{e^{2x}e^{-2x}\left(1 + 2e^x + e^{2x}\right)} = \frac{e^x}{e^{2x}\left(e^{-2x} + 2e^{-x} + 1\right)} = \frac{e^{-x}}{\left(1 + e^{-x}\right)^2} \tag{40}$$

Second

$$\cosh x = \frac{e^x + e^{-x}}{2} \quad \text{and} \quad \cosh^2\left(\tfrac{1}{2}x\right) = \left(\frac{e^{\frac{1}{2}x} + e^{-\frac{1}{2}x}}{2}\right)^2 = \frac{e^x + 2 + e^{-x}}{4}$$

and with $e^x\left(1 + e^{-x}\right)^2 = e^x\left(1 + 2e^{-x} + e^{-2x}\right) = e^x + 2 + e^{-x}$ then $\cosh^2\left(\tfrac{1}{2}x\right) = \dfrac{e^x\left(1 + e^{-x}\right)^2}{4}$ and

$$\tfrac{1}{4}\operatorname{sech}^2\left(\tfrac{1}{2}x\right) = \frac{e^{-x}}{\left(1 + e^{-x}\right)^2} \tag{41}$$

Third

$$\frac{d}{dx}\tanh u = \operatorname{sech}^2 u \frac{du}{dx} \quad \text{so} \quad \tfrac{1}{2}\frac{d}{dt}\left(\tanh\left(\tfrac{1}{2}x\right)\right) = \tfrac{1}{4}\operatorname{sech}^2\left(\tfrac{1}{2}x\right) \tag{42}$$

Now to evaluate (39) it is useful to note that $\displaystyle\int_{-\infty}^{+\infty} \frac{te^{-t}}{\left(1 + e^{-t}\right)^2} dt = 0$ since the integrand $f\left(t\right) = \dfrac{te^{-t}}{\left(1 + e^{-t}\right)^2}$ is an odd function of t since $f\left(-t\right) = -f\left(t\right)$ and the interval of integration is symmetric. Hence the mean becomes

$$\mu_X = a\int_{-\infty}^{+\infty} \frac{e^{-t}}{\left(1 + e^{-t}\right)^2} dt \tag{43}$$

Now, using (40), (41) and (42)

$$\int_{-\infty}^{+\infty} \frac{e^{-t}}{\left(1 + e^{-t}\right)^2} dt = \int_{-\infty}^{+\infty} \tfrac{1}{4}\operatorname{sech}^2\left(\tfrac{1}{2}t\right) dt = \tfrac{1}{2}\int_{-\infty}^{+\infty} \frac{d}{dt}\left(\tanh\left(\tfrac{1}{2}t\right)\right) dt = \tfrac{1}{2}\left[\tanh\left(\tfrac{1}{2}t\right)\right]_{-\infty}^{+\infty} = \tfrac{1}{2}\left[1 - \left(-1\right)\right] = 1 \tag{44}$$

and using this result in (43) gives the mean of the Logistic distribution as

$$\mu_X = a \tag{45}$$

The variance

Using (34), (35) and (37) gives

$$\begin{aligned}\sigma_X^2 &= E\left\{\left(X-\mu_X\right)^2\right\} = E\left\{X^2\right\} - \left(E\left\{X\right\}\right)^2 \\ &= \int_{-\infty}^{+\infty} x^2 f_X\left(x\right)dx - \left(\mu_X\right)^2 \\ &= \int_{-\infty}^{+\infty} \frac{x^2 e^{-\left(\frac{x-a}{b}\right)}}{b\left(1+e^{-\left(\frac{x-a}{b}\right)}\right)^2}dx - a^2\end{aligned} \tag{46}$$

and similarly to the derivation of the mean, the substitution $t = \dfrac{x-a}{b}$ gives $x = tb + a$ and $dx = b\,dt$ giving (46) as

$$\sigma_X^2 = \int_{-\infty}^{+\infty} \frac{\left(tb+a\right)^2 e^{-t}}{\left(1+e^{-t}\right)^2}dt - a^2 = \int_{-\infty}^{+\infty} \frac{\left[\left(tb\right)^2 + 2abt + a^2\right]e^{-t}}{\left(1+e^{-t}\right)^2}dt - a^2$$

and

$$\sigma_X^2 = b^2\int_{-\infty}^{+\infty} \frac{t^2\, e^{-t}}{\left(1+e^{-t}\right)^2}dt + 2ab\int_{-\infty}^{+\infty} \frac{t\, e^{-t}}{\left(1+e^{-t}\right)^2}dt + a^2\int_{-\infty}^{+\infty} \frac{e^{-t}}{\left(1+e^{-t}\right)^2}dt - a^2 \tag{47}$$

Now the second integral of (47) equals zero, since the integrand is an odd function of t. And, using (44), the third integral of (47) equals one giving the variance as

$$\sigma_X^2 = b^2\int_{-\infty}^{+\infty} \frac{t^2\, e^{-t}}{\left(1+e^{-t}\right)^2}dt = 2b^2\int_{0}^{\infty} \frac{t^2\, e^{-t}}{\left(1+e^{-t}\right)^2}dt \tag{48}$$

since the function $f\left(t\right) = \dfrac{t^2\, e^{-t}}{\left(1+e^{-t}\right)^2}$ is symmetric about the $f\left(t\right)$ axis.

Now, using the series expression $\dfrac{1}{\left(1+x\right)^2} = 1 - 2x + 3x^2 - 4x^3 + 5x^4 - \cdots$

$$\frac{e^{-x}}{\left(1+e^{-x}\right)^2} = e^{-x} - 2e^{-2x} + 3e^{-3x} - 4e^{-4x} + 5e^{-5x} - \cdots = \sum_{n=1}^{\infty} n\left(-1\right)^{n-1} e^{-nx} \tag{49}$$

and using (49) in (48) gives

$$\sigma_X^2 = 2b^2\int_0^{\infty} t^2 \sum_{n=1}^{\infty} n\left(-1\right)^{n-1} e^{-nt}dt = 2b^2\sum_{n=1}^{\infty} n\left(-1\right)^{n-1}\int_0^{\infty} t^2 e^{-nt}dt$$

Using the standard integral result $\int x^2 e^{ax}\,dx = \dfrac{e^{ax}}{a}\left(x^2 - \dfrac{2x}{a} + \dfrac{2}{a^2}\right)$ gives

$$
\begin{aligned}
\sigma_X^2 &= 2b^2 \sum_{n=1}^{\infty} n\left(-1\right)^{n-1}\left[\frac{e^{-nt}}{-n}\left(t^2 - \frac{2t}{-n} + \frac{2}{\left(-n\right)^2}\right)\right]_0^{\infty} \\
&= 2b^2 \sum_{n=1}^{\infty} n\left(-1\right)^{n-1}\left[0 - \frac{e^0}{-n}\left(\frac{2}{n^2}\right)\right] \\
&= 2b^2 \sum_{n=1}^{\infty} n\left(-1\right)^{n-1}\frac{2}{n^3} \\
&= 4b^2 \sum_{n=1}^{\infty} \left(-1\right)^{n-1}\frac{1}{n^2}
\end{aligned} \tag{50}
$$

And

$$
\begin{aligned}
\sum_{n=1}^{\infty}\left(-1\right)^{n-1}\frac{1}{n^2} &= 1 - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \frac{1}{5^2} - \cdots \\
&= 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \cdots - 2\left(\frac{1}{2^2} + \frac{1}{4^2} + \frac{1}{6^2} + \cdots\right) \\
&= \sum_{n=1}^{\infty}\frac{1}{n^2} - \frac{2}{2^2}\sum_{n=1}^{\infty}\frac{1}{n^2} \\
&= \frac{\pi^2}{6} - \frac{\pi^2}{12} \\
&= \frac{\pi^2}{12}
\end{aligned} \tag{51}
$$

[Note here Euler's remarkable result: $\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}$ (Euler 1784)]

Hence, using (51) in (50) gives the variance of the Logistic distribution as

$$
\sigma_X^2 = \frac{b^2\pi^2}{3} \tag{52}
$$

The standard deviation σ (the positive square-root of the variance) and the shape parameter b are

$$
\sigma = \frac{b\pi}{\sqrt{3}}, \qquad b = \frac{\sigma\sqrt{3}}{\pi} \tag{53}
$$

If X is a random variable having a Logistic distribution with parameters μ and σ the usual statistical notation is $X \sim LOG\left(\mu,\sigma\right)$ and the probability density function is

$$
f_X\left(x : \mu,\sigma\right) = \frac{\pi}{\sigma\sqrt{3}}\frac{e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)}}{\left(1 + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)}\right)^2} = \frac{\pi}{4\sigma\sqrt{3}}\operatorname{sech}^2\frac{\pi}{2\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right) \tag{54}
$$

where $-\infty < \mu < \infty$ and $\sigma > 0$ are parameters.

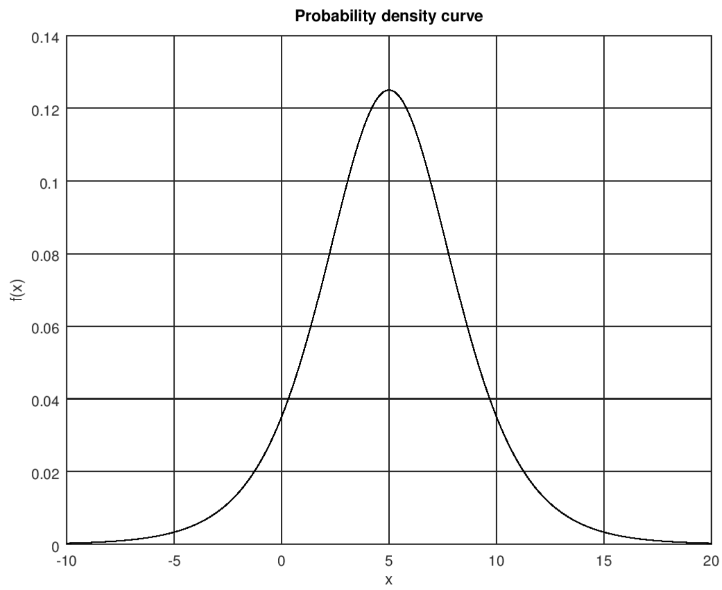

Figure 5. Probability density curve: $f_X\left(x : \mu,\sigma\right)$, $\mu = 5$ and $\sigma = \dfrac{2\pi}{\sqrt{3}}$

The cumulative distribution function of the Logistic distribution is

$$F_X\left(x : \mu,\sigma\right) = \frac{1}{1 + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)}} = \tfrac{1}{2}\left(1 + \tanh\tfrac{1}{2}\left(\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)\right)\right) \tag{55}$$

using $\tanh x = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$ and $\tfrac{1}{2}\left(1 + \tanh x\right) = \dfrac{e^x}{e^x + e^{-x}} = \dfrac{1}{1 + e^{-2x}}$

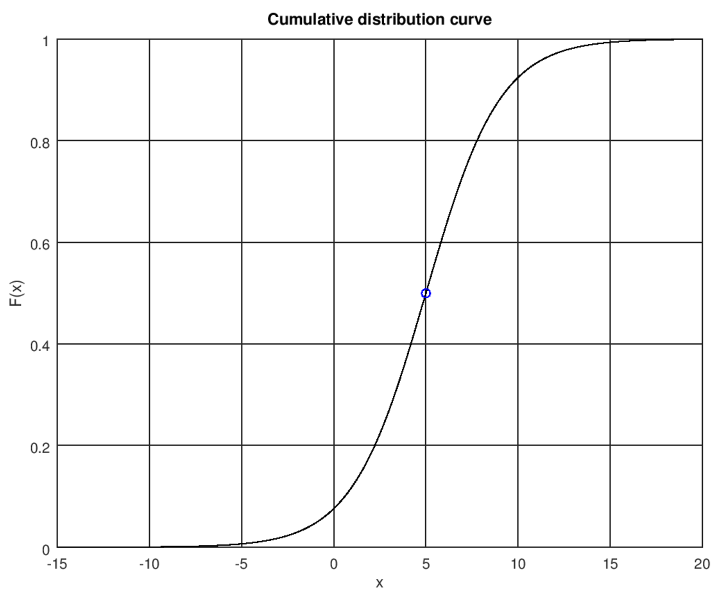

Figure 6. Cumulative distribution curve: $F_X\left(x : \mu,\sigma\right)$, $\mu = 5$ and $\sigma = \dfrac{2\pi}{\sqrt{3}}$

Logistic Regression

Logistic regression was developed by statistician David Cox (Cox 1958) as a means of measuring the relationship between a binary dependent variable (yes or no, win or loss, 1 or 0, etc.) and one or more independent variables by estimating probabilities using a logistic function. The key features of logistic regression are (i) the conditional distribution $y \mid x$ is a Bernoulli distribution[2] where $y \mid x$ means y given x and (ii) the predicted values are probabilities and are therefore restricted to (0,1).

The model for logistic regression is the function

$$y = \frac{1}{1+e^{-z}} \tag{56}$$

where $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$. y is the predicted variable (probability), $x_1, x_2, x_3, \ldots$ are independent variables (parameters) and $\beta_0, \beta_1, \beta_2, \ldots$ are coefficients.

An expression for z may be obtained from (56) as follows

$$\begin{aligned} y &= \frac{1}{1+e^{-z}} \\ y + ye^{-z} &= 1 \\ e^{-z} &= \frac{1-y}{y} \\ e^{z} &= \frac{y}{1-y} \end{aligned}$$

Taking natural logarithms of both sides, noting that $\ln e^{z} = z$ and $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$ gives

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \ln\left(\frac{y}{1-y}\right) \tag{57}$$

This is a non-linear function relating probabilities y with independent variables x and coefficients β.

To enable the determination of coefficients of variables given probabilities it would be desirable to have a linear function relating these quantities and this may be achieved with the aid of Taylor's theorem, that for a single variable, can be expressed in the following form

$$\begin{aligned} f(x) = f(a) + \left.\frac{df}{dx}\right|_a (x-a) + \left.\frac{d^2 f}{dx^2}\right|_a \frac{(x-a)^2}{2!} + \left.\frac{d^3 f}{dx^3}\right|_a \frac{(x-a)^3}{3!} + \cdots \\ + \left.\frac{d^{n-1} f}{dx^{n-1}}\right|_a \frac{(x-a)^{n-1}}{(n-1)!} + R_n \end{aligned} \tag{58}$$

where R_n is the remainder after n terms and $\lim_{n\to\infty} R_n = 0$ for $f(x)$ about $x = a$ and $\left.\frac{df}{dx}\right|_a$, $\left.\frac{d^2 f}{dx^2}\right|_a$, etc. are derivatives of the function $f(x)$ evaluated at $x = a$.

Using Taylor's theorem on the right-hand-side of (57) and evaluating the derivatives about the point $y = p$ gives

[2] The probability distribution (in honour of the Swiss mathematician Jacob Bernoulli) of a random variable which takes the value of 1 with probability p and the value of 0 with the probability $q = 1 - p$.

$$\ln\left(\frac{y}{1-y}\right) = \ln\left(\frac{p}{1-p}\right) + \frac{1}{p(1-p)}(y-p) + \frac{2p-1}{2p^2(1-p)^2}(y-p)^2 + \frac{3p^2-3p+1}{3p^3(1-p)^3}(y-p)^3 + \cdots$$
$$= \ln\left(\frac{p}{1-p}\right) + \frac{1}{p(1-p)}(y-p) + \text{ higher order terms} \tag{59}$$

If p is a close approximation of y the term $(y-p)$ is small, $(y-p)^2$ is very much smaller and $(y-p)^3$ exceedingly small and we may neglect the higher order terms in (59) and write

$$\ln\left(\frac{y}{1-y}\right) \cong \ln\left(\frac{p}{1-p}\right) + \frac{y-p}{p(1-p)}$$

Substituting this result into (57) gives

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \cong \ln\left(\frac{p}{1-p}\right) + \frac{y-p}{p(1-p)} \tag{60}$$

Now, suppose that the approximation p of the probability y is obtained from

$$p = \frac{1}{1+e^{-z^0}} \tag{61}$$

where $z^0 = \beta_0^0 + \beta_1^0 x_1 + \beta_2^0 x_2 + \cdots + \beta_n^0 x_n$ and $\beta_0^0, \beta_1^0, \beta_2^0, \ldots$ are approximate values of the coefficients $\beta_0, \beta_1, \beta_2, \ldots$. And bearing in mind (57)

$$\beta_0^0 + \beta_1^0 x_1 + \beta_2^0 x_2 + \cdots + \beta_n^0 x_n = \ln\left(\frac{p}{1-p}\right) \tag{62}$$

Now, using (62) in (60) and adding a residual v to the left-hand-side to account for small random errors in the variables x, an equation than can be used to determine the coefficients β in an iterative scheme is

$$v + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \beta_0^0 + \beta_1^0 x_1 + \beta_2^0 x_2 + \cdots + \beta_n^0 x_n + \frac{y-p}{p(1-p)} \tag{63}$$

To see how this may work, suppose as an example, 20 students sitting an examination with the observed outcomes $y_1, y_2, \ldots, y_{20}$ as pass or fail and recorded as 1 or 0. These outcomes are thought to be related to a single variable x that is hours of study in preparation for the exam and the logistic function is assumed to be $y = \dfrac{1}{1+e^{-(\beta_0+\beta_1 x)}}$. The wish is to determine the two coefficients β_0 and β_1.

Now assume approximate values β_0^0, β_1^0 for the coefficients and following (63) the observation equation for the k^{th} outcome is

$$v_k + \beta_0 + \beta_1 x_k = \beta_0^0 + \beta_1^0 x_k + \frac{y_k - p_k}{p_k(1-p_k)} \tag{64}$$

and for the 20 observed outcomes the following equations arise

$$
\begin{aligned}
v_1 + \beta_0 + \beta_1 x_1 &= \beta_0^0 + \beta_1^0 x_1 + \frac{y_1 - p_1}{p_1\left(1 - p_1\right)} \\
v_2 + \beta_0 + \beta_1 x_2 &= \beta_0^0 + \beta_1^0 x_2 + \frac{y_2 - p_2}{p_2\left(1 - p_2\right)} \\
v_3 + \beta_0 + \beta_1 x_3 &= \beta_0^0 + \beta_1^0 x_3 + \frac{y_3 - p_3}{p_3\left(1 - p_3\right)} \\
&\vdots \\
v_{20} + \beta_0 + \beta_1 x_{20} &= \beta_0^0 + \beta_1^0 x_{20} + \frac{y_{20} - p_{20}}{p_{20}\left(1 - p_{20}\right)}
\end{aligned}
$$

These equations can be rearranged into matrix form

$$
\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_{20} \end{bmatrix} + \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_{20} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_{20} \end{bmatrix} \begin{bmatrix} \beta_0^0 \\ \beta_1^0 \end{bmatrix} + \begin{bmatrix} \left(y_1 - p_1\right)\big/\left(p_1\left(1 - p_1\right)\right) \\ \left(y_2 - p_2\right)\big/\left(p_2\left(1 - p_2\right)\right) \\ \left(y_3 - p_3\right)\big/\left(p_3\left(1 - p_3\right)\right) \\ \vdots \\ \left(y_{20} - p_{20}\right)\big/\left(p_{20}\left(1 - p_{20}\right)\right) \end{bmatrix} \tag{65}
$$

or

$$
\mathbf{v} + \mathbf{B}\mathbf{x} = \mathbf{f}
$$

$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_{20} \end{bmatrix}$ is a vector of residuals, $\mathbf{B} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_{20} \end{bmatrix}$ is a coefficient matrix, $\mathbf{x} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ is the vector of coefficients

and $\mathbf{f} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_{20} \end{bmatrix} \begin{bmatrix} \beta_0^0 \\ \beta_1^0 \end{bmatrix} + \begin{bmatrix} \left(y_1 - p_1\right)\big/\left(p_1\left(1 - p_1\right)\right) \\ \left(y_2 - p_2\right)\big/\left(p_2\left(1 - p_2\right)\right) \\ \left(y_3 - p_3\right)\big/\left(p_3\left(1 - p_3\right)\right) \\ \vdots \\ \left(y_{20} - p_{20}\right)\big/\left(p_{20}\left(1 - p_{20}\right)\right) \end{bmatrix}$ is a vector of numeric terms that can be rearranged as

$$
\mathbf{f} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_{20} \end{bmatrix} \begin{bmatrix} \beta_0^0 \\ \beta_1^0 \end{bmatrix} - \begin{bmatrix} 1\big/\left(1 - p_1\right) \\ 1\big/\left(1 - p_2\right) \\ 1\big/\left(1 - p_3\right) \\ \vdots \\ 1\big/\left(1 - p_{20}\right) \end{bmatrix} + \begin{bmatrix} y_1\big/\left(p_1\left(1 - p_1\right)\right) \\ y_2\big/\left(p_2\left(1 - p_2\right)\right) \\ y_3\big/\left(p_3\left(1 - p_3\right)\right) \\ \vdots \\ y_{20}\big/\left(p_{20}\left(1 - p_{20}\right)\right) \end{bmatrix} = \mathbf{B}\mathbf{x}^0 - \mathbf{c} + \mathbf{A}\mathbf{y} = \mathbf{d} + \mathbf{A}\mathbf{y} \tag{66}
$$

where $\mathbf{x}^0 = \begin{bmatrix} \beta_0^0 \\ \beta_1^0 \end{bmatrix}$ is the vector of approximate coefficients, $\mathbf{c} = \begin{bmatrix} 1\big/\left(1 - p_1\right) \\ 1\big/\left(1 - p_2\right) \\ \vdots \\ 1\big/\left(1 - p_{20}\right) \end{bmatrix}$ is a vector of numeric terms,

$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{20} \end{bmatrix}$ is the vector of measurements,

$$\mathbf{A} = \begin{bmatrix} \dfrac{1}{p_1\left(1-p_1\right)} & 0 & \cdots & 0 \\ 0 & \dfrac{1}{p_2\left(1-p_2\right)} & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{1}{p_{20}\left(1-p_{20}\right)} \end{bmatrix}$$ is a diagonal coefficient matrix and

$\mathbf{d} = \mathbf{B}\mathbf{x}^0 - \mathbf{c}$ is a vector of numeric terms.

Applying propagation of variances[3] to (66) gives

$$\mathbf{V}_{ff} = \mathbf{A}\mathbf{V}_{yy}\mathbf{A}^T \tag{67}$$

where $\mathbf{V}_{yy}$ is a diagonal matrix containing the variances of the measurements y and $\mathbf{V}_{ff}$ is a diagonal matrix containing the variances of the numeric terms **f** in (66). The *y's* in the right-hand-side of (66) are random variables that follow a Bernoulli distribution and take the value of 1 with a probability of p and a value of 0 with a probability of $q = 1 - p$. The variance of these variables is $p\left(1-p\right)$ and the general form of $\mathbf{V}_{ff}$ is given by

$$\mathbf{V}_{ff} = \begin{bmatrix} \dfrac{1}{p_1\left(1-p_1\right)} & 0 & \cdots & 0 \\ 0 & \dfrac{1}{p_2\left(1-p_2\right)} & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{1}{p_{20}\left(1-p_{20}\right)} \end{bmatrix} \tag{68}$$

Now, with the general relationship that weights are inversely proportional to variances, i.e. $\mathbf{W} = \mathbf{V}^{-1}$ the general form of the weight matrix **W** of (65) is

$$\mathbf{W} = \begin{bmatrix} p_1\left(1-p_1\right) & 0 & \cdots & 0 \\ 0 & p_2\left(1-p_2\right) & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & p_{20}\left(1-p_{20}\right) \end{bmatrix} \tag{69}$$

The coefficients in the vector **x** in (65) can now be solved using *least squares*[4] with the standard result (Mikhail 1976)

$$\mathbf{x} = \left(\mathbf{B}^T\mathbf{W}\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{W}\mathbf{f} \tag{70}$$

and these are 'updates' of the approximate values in $\mathbf{x}^0$. This iterative process can be repeated until there is no appreciable change in the updated coefficients. In the literature associated with logistic regression this iterative least squares process of determining the coefficients is known as *Iteratively Reweighted Least Squares* (IRLS).

[3] Propagation of Variances is a mathematical technique of estimating the variance of functions of random variables that have assumed (or known) variances. See Appendix A.

[4] Least squares is a mathematical estimation process used to calculate the best estimate of quantities from overdetermined systems of equations. It is commonly used to determine the line of best fit through a number of data points and this application is known as linear regression.

A MATLAB[5] function *logistic.m* that solves for the parameters of a logistic regression has been developed by Professor Geoffrey J. Gordon in the Machine Learning Department at Carnegie Mellon University, USA and is available from his website: http://www.cs.cmu.edu/~ggordon/IRLS-example/. This function will also run under GNU OCTAVE[6]. A copy of the function and an example of its use is shown in Appendix B.

The data for the example used for explanation (20 students studying for an examination) is from the Wikipedia page *Logistic Regression* (https://en.wikipedia.org/wiki/Logistic_regression) and is shown in the table below.

	Hours	Pass		Hours	Pass		Hours	Pass		Hours	Pass
1	0.50	0	6	1.75	0	11	2.75	1	16	4.25	1
2	0.75	0	7	1.75	1	12	3.00	0	17	4.50	1
3	1.00	0	8	2.00	0	13	3.25	1	18	4.75	1
4	1.25	0	9	2.25	1	14	3.50	0	19	5.00	1
5	1.50	0	10	2.50	0	15	4.00	1	20	5.5	1

Table 1. Number of hours each student spent studying, and whether they passed (1) of failed (0)

20 students sit an examination with the observed outcomes $y_1, y_2, \ldots, y_{20}$ as pass or fail and recorded as 1 or 0. These outcomes are related to a single variable x that is hours of study in preparation for the exam and the logistic function is assumed to be $y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$. The coefficients β_0 and β_1 are computed using the MATLAB function *logistic.m.* The input data and results are shown below

```
>> a
a =

   1.00000   0.50000
   1.00000   0.75000
   1.00000   1.00000
   1.00000   1.25000
   1.00000   1.50000
   1.00000   1.75000
   1.00000   1.75000
   1.00000   2.00000
   1.00000   2.25000
   1.00000   2.50000
   1.00000   2.75000
   1.00000   3.00000
   1.00000   3.25000
   1.00000   3.50000
   1.00000   4.00000
   1.00000   4.25000
   1.00000   4.50000
   1.00000   4.75000
   1.00000   5.00000
   1.00000   5.50000

>> y'
ans =

   0   0   0   0   0   0   1   0   1   0   1   0   1   0   1   1   1   1   1   1
```

[5] MATLAB (*matrix laboratory*) is a numerical computing environment and programming language developed by MathWorks.

[6] GNU OCTAVE is free software featuring a high-level programming language, primarily intended for numerical computations that is mostly compatible with MATLAB. It is part of the GNU Project and is free software under the terms of the GNU General Public License.

```
>> xhat = logistic(a, y, [], [], struct('verbose', 1))
  1: [ -2.61571 0.938375 ]
  2: [ -3.66126 1.33855 ]
  3: [ -4.03513 1.48734 ]
  4: [ -4.07709 1.5044 ]
  5: [ -4.07757 1.5046 ]
  6: [ -4.07757 1.5046 ]
  7: [ -4.07757 1.5046 ]
  8: [ -4.07757 1.5046 ]
Converged.
xhat =

  -4.0776
   1.5046
```

The array `a` is the coefficient matrix $\mathbf{B}$ and the vector `y` is $\mathbf{y}$ noting that $y' = \mathbf{y}^T$

The iterative process has converged after 8 iterations and the coefficients are in the array `xhat` where $\beta_0 = -4.0776$ and $\beta_1 = 1.5046$.

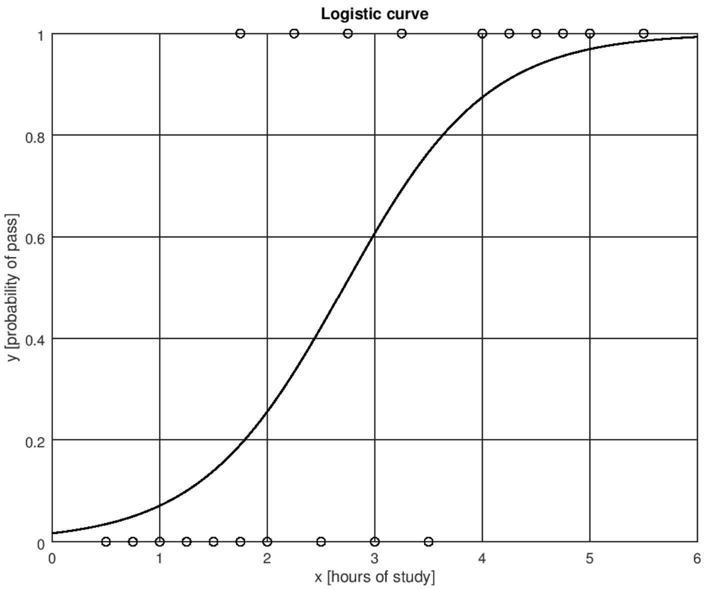

Figure 7. Probability y of passing exam after x hours of study.

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \text{ where } \beta_0 = -4.0776 \text{ and } \beta_1 = 1.5046$$

The midpoint of the symmetric curve is the point $\left(x_0, \frac{1}{2}\right)$, and $y = \frac{1}{2}$ when the exponent is zero, i.e., when

$\beta_0 + \beta_1 x_0 = 0$ giving $x_0 = -\dfrac{\beta_0}{\beta_1} = 2.7101$

In this example, a student who studies 2 hours has a probability of passing the exam of 0.26

$$P\left(\text{pass}\right) = \frac{1}{1 + e^{-\left(-4.0776 + 1.5046 \times 2\right)}} = 0.2557$$

and similarly, a student who studies 4 hours has a probability of passing the exam of 0.87

$$P\left(\text{pass}\right) = \frac{1}{1 + e^{-\left(-4.0776 + 1.5046 \times 4\right)}} = 0.8744$$

Table 2 shows probabilities of passing the exam for several values of hours of study

Hours of study	**Probability of passing exam**
1	0.07
2	0.26
2.7101	0.50
3	0.61
4	0.87
5	0.97

Table 2.

Logistic Curve for Dove Open Doubles Petanque Tournament 01-May-2016

In this tournament there were 18 teams. There were 4 Qualifying rounds (Swiss System[7]) and the top 9 teams were seeded and did not play each other in Round 1 of the Qualifying. After the Qualifying the teams were ranked 1–18 and teams ranked 1 to 8 went into a Principale and teams ranked 9 to 16 went into a Complémentaire. The remaining two teams took no further part in the tournament. The Principale and Complémentaire were single elimination finals series with play-off's for 3rd and 4th places. There were 36 matches in the Qualifying and 8 matches each in the Principale and Complémentaire making a total of 52 matches involving the 18 teams.

Tables 3 (Qualifying), 5 (Principale) and 6 (Complémentaire) show matches and game scores in the tournament and Tables 4 and 7 show ranking after Qualifying and final ranking in Principale and Complémentaire respectively.

Round 1			**Round 2**			**Round 3**			**Round 4**		
16(13)	v	**9**(4)	**2**(7)	v	**14**(11)	**3**(13)	v	**14**(8)	**3**(13)	v	**16**(8)
18(6)	v	**14**(10)	**7**(9)	v	**15**(13)	**15**(11)	v	**16**(12)	**18**(12)	v	**7**(13)
5(12)	v	**11**(8)	**8**(3)	v	**16**(13)	**17**(11)	v	**6**(12)	**14**(5)	v	**4**(10)
7(13)	v	**1**(6)	**5**(5)	v	**17**(13)	**5**(10)	v	**18**(13)	**15**(13)	v	**6**(11)
2(13)	v	**12**(0)	**3**(13)	v	**6**(3)	**4**(13)	v	**13**(12)	**1**(12)	v	**13**(9)
8(13)	v	**4**(8)	**4**(10)	v	**11**(6)	**8**(8)	v	**1**(9)	**5**(8)	v	**2**(13)
15(13)	v	**17**(1)	**9**(2)	v	**18**(13)	**2**(10)	v	**7**(13)	**17**(12)	v	**8**(11)
13(10)	v	**6**(13)	**13**(13)	v	**12**(2)	**11**(13)	v	**10**(4)	**11**(11)	v	**12**(8)
3(13)	v	**10**(7)	**10**(3)	v	**1**(13)	**9**(8)	v	**12**(9)	**10**(9)	v	**9**(11)

Table 3. Dove Open Doubles: Qualifying matches
(game scores shown in parentheses beside team number)

[7] The Swiss System (also known as the Swiss Ladder System) is a tournament system that allows participants to play a limited number of rounds against opponents of similar strength. The system was introduced in 1895 by Dr. J. Muller in a chess tournament in Zurich, hence the name 'Swiss System'. The principles of the system are: (i) In every round, each player is paired with an opponent with an equal score (or as nearly equal as possible); (ii) Two players are paired at most once; (iii) After a predetermined number of rounds the players are ranked according to a set of criteria. The leading player wins; or the ranking is the basis of subsequent elimination series.

Qualifying Ranking							
Rank	**Team**	**Score**	**BHN**	**fBHN**	**Games**	**Points**	**delta**
1	**3**	4	7	40	4:0	52:26	+26
2	**15**	3	10	36	3:1	50:33	+17
3	**7**	3	10	29	3:1	48:41	+7
4	**16**	3	9	34	3:1	46:31	+15
5	**4**	3	6	36	3:1	41:36	+5
6	**1**	3	5	40	3:1	40:33	+7
7	**14**	2	11	27	2:2	34:36	–2
8	**6**	2	10	33	2:2	39:47	–8
9	**17**	2	7	39	2:2	37:41	–4
10	**2**	2	7	35	2:2	43:32	+11
11	**18**	2	7	35	2:2	44:35	+9
12	**11**	2	5	30	2:2	38:34	+4
13	**8**	1	11	27	1:3	35:42	–7
14	**13**	1	9	27	1:3	44:40	+4
15	**5**	1	8	26	1:3	35:47	–12
16	**9**	1	6	32	1:3	25:44	–19
17	**12**	1	6	27	1:3	19:45	–26
18	**10**	0	10	23	0:4	23:50	–27

Table 4. Dove Open Doubles: Ranking after Qualifying rounds
(BHN is Buchholtz Number[8], fBHN is Fine Buchholtz Number)

Quarter-Finals	Semi-Finals	Final	Playoff
3(11) v **6**(9) **16**(11) v **4**(10) **1**(9) v **7**(11) **14**(11) v **15**(9)	**3**(13) v **16**(2) **7**(8) v **14**(9)	**3**(13) v **14**(7)	**16**(6) v **7**(13)

Table 5. Dove Open Doubles: Principale matches

Quarter-Finals	Semi-Finals	Final	Playoff
17(10) v **9**(8) **11**(13) v **8**(4) **13**(2) v **18**(13) **5**(7) v **2**(13)	**17**(2) v **11**(13) **18**(13) v **2**(1)	**11**(10) v **18**(13)	**17**(3) v **2**(13)

Table 6. Dove Open Doubles: Complémentaire matches

Principale		Complémentaire	
Rank	**Team**	**Rank**	**Team**
1	**3**	1	**18**
2	**14**	2	**11**
3	**7**	3	**2**
4	**16**	4	**17**
= 5	**4** **6** **15** **1**	= 5	**8** **13** **9** **5**

Table 7. Dove Open Doubles: Final Ranking in Principale & Complémentaire

[8] The Buchholtz system is a ranking system, first used by Bruno Buchholtz in a Swiss System chess tournament in 1932. The principle of the system is that when two players have equal scores at the end of a defined number of rounds a tie break is required to determine the top ranked player. The scores of both player's opponents (in all rounds) are added giving each their Buchholtz Number (BHN). The player having the larger BHN is ranked higher on the assumption they have played against better performing players. The Fine Buchholtz Number (fBHN) is the sum of the opponents Buchholtz Numbers and is used to break ties where player's BHN are equal. In the rare case that Score, BHN and fBHN are all equal then delta = points For – points Against is used as a tie break (see Teams 2 & 18 in Table 7).

A least squares solution for team ratings, based on the 36 Qualifying matches yielded the following values with the highest rating team, (team 15) having a rating $r_{15} = 100$.

$$\text{ratings} = \left\{ \begin{array}{lll} r_1 = 93.670 & r_7 = 97.207 & r_{13} = 92.312 \\ r_2 = 94.100 & r_8 = 92.923 & r_{14} = 94.837 \\ r_3 = 99.666 & r_9 = 85.586 & r_{15} = 100.000 \\ r_4 = 93.526 & r_{10} = 85.238 & r_{16} = 98.294 \\ r_5 = 89.598 & r_{11} = 89.030 & r_{17} = 93.204 \\ r_6 = 94.295 & r_{12} = 83.757 & r_{18} = 94.057 \end{array} \right\}$$

Using these ratings an analysis of the Qualifying matches (Table 3) yields the following tabulated results where Team A is the first-named team and Team B is the second-named team. A win is recorded as 1 and a loss is recorded as 0, and *dr* is the difference in ratings. If *dr* is positive then Team A is the higher ranking team.

Match	Team A	Team B	$dr = r_A - r_B$	Match	Team A	Team B	$dr = r_A - r_B$
1	1	0	12.708	19	1	0	4.829
2	0	1	-0.780	20	0	1	1.706
3	1	0	0.568	21	0	1	-1.091
4	1	0	3.537	22	0	1	-4.459
5	1	0	10.343	23	1	0	1.214
6	1	0	-0.603	24	0	1	-0.747
7	1	0	6.796	25	0	1	-3.107
8	0	1	-1.983	26	1	0	3.792
9	1	0	14.428	27	0	1	1.829
10	0	1	-0.737	28	1	0	1.372
11	0	1	-2.793	29	0	1	-3.150
12	0	1	-5.371	30	0	1	1.311
13	0	1	-3.606	31	1	0	5.705
14	1	0	5.371	32	1	0	1.358
15	1	0	4.496	33	0	1	-4.502
16	0	1	-8.471	34	1	0	0.281
17	1	0	8.555	35	1	0	5.273
18	0	1	-8.432	36	0	1	-0.348

Table 8. Win/Loss and rating difference for Qualifying matches of Dove Open Doubles.

Using the values in Table 8, Logistic Regression is used to compute the parameters of a curve representing the probability of Team A winning given a certain rating difference.

The curve is assumed to have the following form: $y = \dfrac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$ where y is the probability of winning and x is the rating difference between the two teams and the MATLAB function *logistic.m* is used to compute the coefficients β_0 and β_1.

The input arrays `a` and `y` are shown below together with the output vector `xhat` (Note that `a′` denotes transpose)

```
>> a'
ans =

 Columns 1 through 14:

    1.00000    1.00000    1.00000    1.00000    1.00000    1.00000    1.00000    1.00000
    1.00000    1.00000    1.00000    1.00000    1.00000    1.00000
   12.70800   -0.78000    0.56800    3.53700   10.34300   -0.60300    6.79600   -1.98300
   14.42800   -0.73700   -2.79300   -5.37100   -3.60600    5.37100

Columns 15 through 28:

    1.00000    1.00000    1.00000    1.00000    1.00000    1.00000    1.00000    1.00000
    1.00000    1.00000    1.00000    1.00000    1.00000    1.00000
    4.49600   -8.47100    8.55500   -8.43200    4.82900    1.70600   -1.09100   -4.45900
    1.21400   -0.74700   -3.10700    3.79200    1.82900    1.37200

 Columns 29 through 36:

    1.00000    1.00000    1.00000    1.00000    1.00000    1.00000    1.00000    1.00000
   -3.15000    1.31100    5.70500    1.35800   -4.50200    0.28100    5.27300   -0.34800

>> y'
ans =

   1   0   1   1   1   1   1   0   1   0   0   0   0   1   1   0   1   0   1   0   0   0
   1   0   0   1   0   1   0   0   1   1   0   1   1   0

>>

>> xhat = logistic(a, y, [], [], struct('verbose',1))
  1: [ -0.348917 0.277334 ]
  2: [ -0.485595 0.504256 ]
  3: [ -0.60651 0.733499 ]
  4: [ -0.690405 0.902133 ]
  5: [ -0.722152 0.964799 ]
  6: [ -0.725415 0.970895 ]
  7: [ -0.725443 0.970944 ]
  8: [ -0.725443 0.970944 ]
  9: [ -0.725443 0.970944 ]
 10: [ -0.725443 0.970944 ]
Converged.
xhat =

  -0.72544
   0.97094

>>
```

The iterative process has converged after 10 iterations and the coefficients are in the array `xhat` where $\beta_0 = -0.72544$ and $\beta_1 = 0.97094$.

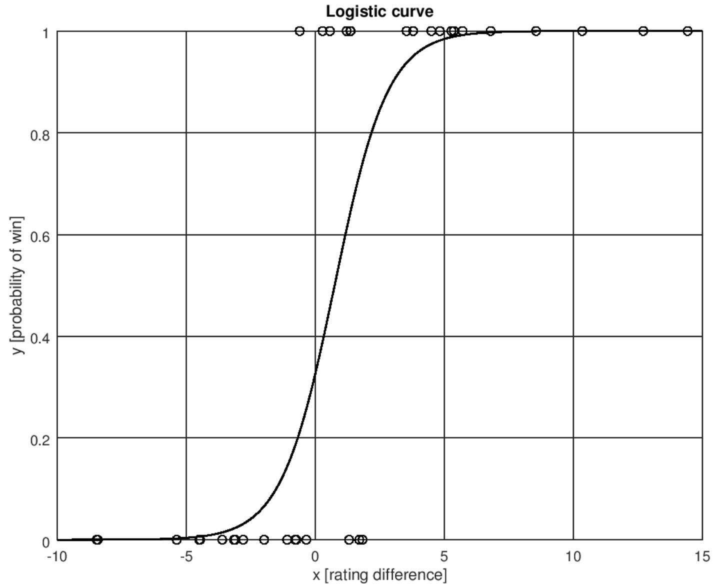

Figure 7. Probability y of winning a petanque match given a rating difference x.

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \text{ where } \beta_0 = -0.72544 \text{ and } \beta_1 = 0.97094$$

The midpoint of the symmetric curve is the point $\left(x_0, \frac{1}{2}\right)$, and $y = \frac{1}{2}$ when the exponent is zero, i.e., when

$\beta_0 + \beta_1 x_0 = 0$ giving $x_0 = -\dfrac{\beta_0}{\beta_1} = 0.74715$

Logistic Curve for Elo Rating System

The Elo Rating System (Elo 1978) is a mathematical process based on a statistical model relating match results to underlying variables representing the abilities of a team or player. The name "Elo" derives from Arpad Elo[9], the inventor of a system for rating chess players and his system, in various modified forms, is used for player or team ratings in many sports.

The Elo Rating System calculates, for every player or team, a numerical rating based on performance in competitions. A rating is a number (usually an integer) between 0 and 3000 that changes over time depending on the outcome of tournament games. The system depends on a curve defined by a *logistic function* (Langville & Meyer 2012, Glickman & Jones, 1999)

$$p_A = \frac{1}{1 + 10^{-\left(\frac{r_A - r_B}{b}\right)}} \tag{71}$$

where p_A is the probability of player A winning in a match A versus B given the player ratings r_A, r_B and b is a shape parameter.

The curve of this function has a similar form to the cumulative distribution curve of the Logistic distribution (33) with e replaced by 10 as a base, $x = r_A - r_B$ is the rating difference between players A and B, $a = 0$ b is the shape parameter.

[9] Arpad Elo (1903 – 1992) the Hungarian-born US physics professor and chess-master who devised a system to rate chess players that was implemented by the United States Chess Federation (USFC) in 1960 and adopted by the World Chess Federation (FIDE) in 1970. Elo described his work in his book *The Rating of Chess Players, Past & Present,* published in 1978 and his system has been adapted to many sports.

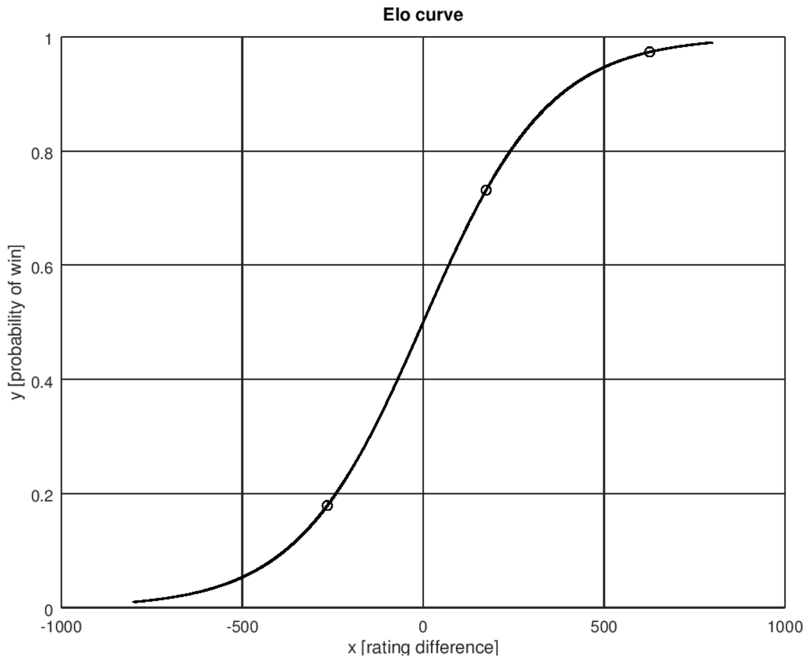

Figure 1. Elo curve: $y = \dfrac{1}{1+10^{-\left(\frac{x}{400}\right)}}$. $y = p_A$ is the probability of A winning, $x = r_A - r_B$ is the rating difference and the shape parameter $b = 400$. The three points on the curve shown thus ○ have rating differences $-265, 174$ and 626 that correspond with probabilities 0.179, 0.731 and 0.973 respectively.

The curve in Figure 3 has the shape parameter $b = 400$ and this value is chosen so that a player rating difference of approximately 200 corresponds to a probability of winning of approximately 0.75. With $y = p_A$ and $x = r_A - r_B$ (71) becomes

$$y = \frac{1}{1+10^{-\left(\frac{x}{400}\right)}} \tag{72}$$

and this equation can be rearranged as $10^{-\left(\frac{x}{400}\right)} = \dfrac{1-y}{y}$. Now using the rule for logarithms that if $p = \log_a N$ then $N = a^p$ the expression for rating difference x is

$$x = -400\log_{10}\left(\frac{1-y}{y}\right) \tag{73}$$

And if probability $y = 0.75$ then rating difference $x = 190.848501$.

For example, suppose two players A and B with ratings 1862 and 1671 respectively play a match. The probability of A winning is given by (71) as

$$p_A = \frac{1}{1+10^{-\left(\frac{1862-1671}{400}\right)}} = \frac{1}{1+10^{-\left(\frac{191}{400}\right)}} = \frac{1}{1+10^{-0.4775}} = 0.750163482$$

We might express this probability of A winning as:

(i) If A played B in 100 matches then A would win 75 of them (75.0163482 actually), or

(ii) A has a 75% chance of winning.

Elo's logistic function uses exponents with a base of 10. This is the base of common logarithms and the following relationships may be useful.

If
$$y = \frac{1}{1+e^{-x}} = \frac{1}{1+10^{-\alpha x}} \tag{74}$$

noting that $e^{x} = 10^{\alpha x}$ and $e^{1} = 2.718281828459... = 10^{\alpha}$

then
$$\alpha = \log_{10} e = 0.434294481903... = \frac{1}{2.302585092994...} \tag{75}$$

Alternatively,

if
$$y = \frac{1}{1+10^{-x}} = \frac{1}{1+e^{-\beta x}} \tag{76}$$

noting that $10^{x} = e^{\beta x}$ and $10^{1} = 10 = e^{\beta}$

then
$$\beta = \ln 10 = 2.302585092994... = \frac{1}{0.434294481903...}$$

References

Bacaër, N, 2011, *A Short History of Mathematical Population Dynamics,* Chapter 6, Verhulst and the logistic equation (1838), pp.35-39, Springer-Verlag London Limited.
http://webpages.fc.ul.pt/~mcgomes/aulas/dinpop/Mod13/Verhulst.pdf [accessed 10-May-2018]

Cox, D.R., 1958, 'The regression analysis of binary sequences', *Journal of the Royal Statistical Society. Series B (Methodological),* Vol. 20, No. 2, pp. 215-242.
http://www.jstor.org/stable/2983890 [accessed 10-May-2018]

Cramer, J.S., 2002, *The Origins of Logistic Regression,* Tinbergen Institute Discussion Paper, TI 2002-119/4, Faculty of Economics and Econometrics, University of Amsterdam, and Timbergen Institute, 14 pages, November 2002.
https://papers.tinbergen.nl/02119.pdf [accessed 15-May-2018]

Elo, A.E., 1978, *The Rating of Chess Players, Past & Present,* 2nd printing, April 2008, Ishi Press International.

Euler, L., 1748, *Introduction to Analysis of the Infinite* (On the use of the Discovered Fractions to Sum Infinite Series), in *Pi: A Source Book,* by Berggren, L., Borwein, J. and Borwein, P., 1997, Springer, New York.

Glickman, M.E. and Jones, A., 1999, 'Rating the chess rating system', *Chance,* Vol. 12, No. 2, pp. 21-28.
http://glicko.net/research/chance.pdf [accessed 25-Aug-2018]

Hunter, M.N, 2018, 'Some comments about The Logistic Function', Private correspondence, 6 pages, 01-Sep-2018.

Johnson, N.L. and Leone, F.C., 1964, *Statistics and Experimental Design In Engineering and the Physical Sciences,* Vol. I, John Wiley & Sons, Inc., New York

Kreyszig, Erwin, 1970, *Introductory Mathematical Statistics,* John Wiley & Sons, New York.

Langville, A.N. and Meyer, C.D., 2012, *Who's #1? The Science of Rating and Ranking,* Princeton University Press, Princeton.

Mikhail, E.M., 1976, *Observations and Least Squares,* IEP—A Dun-Donnelley, New York

O'Connor, J.J. and Robertson, E.F., 2014, 'Pierre François Verhulst', *MacTutor History of Mathermatics,*
http://www-history.mcs.st-andrews.ac.uk/Biographies/Verhulst.html [accessed 16-May-2018]

Verhulst, P.F., 1838, 'Notice sur la loi que la population suit dans son accroissement', *Correspondance Mathématique et Physique,* Publiée par A. Quetelet, Vol. 4, pp. 113-121
https://books.google.com.au/books?id=8GsEAAAAYAAJ&hl=fr&pg=PA113&redir_esc=y#v=onepage&q&f=false [accessed 15-May-2018]

Verhulst, P.F., 1844, 'Recherches mathématiques sur la loi d'accroissement de la population', *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles,* Vol. 18, pp. 1-38
http://www.med.mcgill.ca/epidemiology/Hanley/anniversaries/ByTopic/Verhulst1844.pdf [accessed 15-May-2018]

APPENDIX A: Some Statistical Definitions

Experiments, Sets, Sample Spaces, Events and Probability

The term *statistical experiment* can be used to describe any process by which several chance observations are obtained. All possible outcomes of an experiment comprise a *set* called the *sample space* and a set or sample space contains N elements or members. An *event* is a subset of the sample space containing n elements. Experiments, sets, sample spaces and events are the fundamental tools used to determine the *probability* of certain events where probability is defined as

$$P\left(Event\right) = \frac{n}{N} \tag{77}$$

For example, if a card is drawn from a deck of playing cards, what is the probability that it is a heart? In this case, the experiment is the drawing of the card and the possible outcomes of the experiment could be one of 52 different cards, i.e., the sample space is the set of $N = 52$ possible outcomes and the event is the subset containing $n = 13$ hearts. The probability of drawing a heart is

$$P\left(\text{Heart}\right) = \frac{n}{N} = \frac{13}{52} = 0.25$$

This definition of probability is a simplification of a more general concept of probability that can be explained in the following manner (see Johnson & Leone, 1964, pp.32-3).

> Suppose observations are made on a series of occasions (often termed trials) and during these trials it is noted whether or not a certain event occurs. The event can be almost any observable phenomenon, for example, that the height of a person walking through a doorway is greater than 1.8 metres, that a family leaving a cinema contains three children, that a defective item is selected from an assembly line, and so on. These trials could be conducted twice a week for a month, three times a day for six months or every hour for every day for 10 years. In the theoretical limit, the number of trials N would approach infinity and we could assume, at this point, that we had noted every possible outcome. Therefore, as $N \to \infty$ then N becomes the number of elements in the sample space containing all possible outcomes of the trials. Now for each trial we note whether or not a certain event occurs, so that at the end of N trials we have noted n_N events. The probability of the event (if it in fact occurs) can then be defined as
>
> $$P\left(Event\right) = \lim_{N\to\infty}\left(\frac{n_N}{N}\right)$$
>
> Since n_N and N are both non-negative numbers and n_N is not greater than N then
>
> $$0 \le \frac{n_N}{N} \le 1$$
>
> Hence
>
> $$0 \le P\left\{Event\right\} \le 1$$
>
> If the event occurs at every trial then $n_N = N$ and $n_N / N = 1$ for all N and so $P\left(Event\right) = 1$. This relationship can be described as: *the probability of a certain (or sure) event is equal to 1.*
>
> If the event never occurs, then $n_N = 0$ and $n_N / N = 0$ for all N and so $P\left(Event\right) = 0$. This relationship can be described as: *the probability of an impossible event is zero.*
>
> The converse of these two relationships need not hold, i.e., a probability of one need not imply certainty since it is possible that $\lim_{N\to\infty} n_N / N = 1$ without $n_N = 1$ for all values of N and a

probability of zero need not imply impossibility since it is possible that $\lim_{N\to\infty} n_N \big/ N = 0$ even though $n_N > 0$. Despite these qualifications, it is useful to think of probability as measured on a scale varying from (near) impossibility at 0 to (near) certainty at 1. It should also be noted that this definition of probability (or any other definition) is not directly verifiable in the sense that we cannot actually carry out the infinite series of trials to see whether there really is a unique limiting value for the ratio $n_N \big/ N$. The justification for this definition of probability is utilitarian, in that the results of applying theory based on this definition prove to be useful and that it fits with intuitive ideas. However, it should be realized that it is based on the *concept* of an infinitely long series of trials rather than an actual *series,* however long it may be.

Random Variables and Probability Distributions of Random Variables

A *random variable* X is a rule or a function, which associates a real number with each point in a sample space. As an example, consider the following experiment where two identical coins are tossed; h denotes a head and t denotes a tail.

Experiment: Toss two identical coins.

Sample space: $S = \left\{hh, ht, th, tt\right\}$.

Random Variable: X, the number of heads obtained, may be written as

$$\begin{aligned} X\left(hh\right) &= 2 \\ X\left(ht\right) &= 1 \\ X\left(th\right) &= 1 \\ X\left(tt\right) &= 0 \end{aligned}$$

In this example X is the random variable defined by the rule: "the number of heads obtained". The possible values (or real numbers) that X may take are 0, 1, 2. These possible values are usually denoted by x and the notation $X = x$ denotes x as a possible real value of the random variable X.

Random variables may be discrete or continuous. A *discrete random variable* assumes each of its possible values with a certain probability. For example, in the experiment above; the tossing of two coins, the sample space $S = \left\{hh, ht, th, tt\right\}$ has $N = 4$ elements and the probability the random variable X (the number of heads) assumes the possible values 0, 1 and 2 is given by

x	0	1	2
$P\left(X = x\right)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

Note that the values of x exhaust all possible cases and hence the probabilities add to 1

A *continuous random variable* has a probability of zero of assuming any of its values and consequently, its probability distribution cannot be given in tabular form. The concept of the probability of a continuous random variable assuming a particular value equals zero may seem strange, but the following example illustrates the point. Consider a random variable whose values are the heights of all people over 21 years of age. Between any two values, say 1.75 metres and 1.85 metres, there are an infinite number of heights, one of which is 1.80 metres. The probability of selecting a person at random exactly 1.80 metres tall and not one of the infinitely large set of heights so close to 1.80 metres that you cannot humanly measure the difference is extremely remote, and thus we assign a probability of zero to the event. It follows that probabilities of continuous random variables are defined by specifying an interval within which the random variable lies and it does not matter whether an end-point is included in the interval or not.

$$\begin{aligned} P\left(a < X \le b\right) &= P\left(a < X < b\right) + P\left(X = b\right) \\ &= P\left(a < X < b\right) \end{aligned}$$

It is most convenient to represent all the probabilities of a random variable X by a formula or function denoted by $f_X(x)$, $g_X(x)$, $h_X(x)$, etc., or by $F_X(x)$, $G_X(x)$, $H_X(x)$, etc.

In this notation the subscript X denotes that $f_X(x)$ or $F_X(x)$ is a function of the random variable X which takes the numerical values x within the function. Such functions are known as *probability distribution functions* and they are paired; i.e., $f_X(x)$ pairs with $F_X(x)$, $g_X(x)$ pairs with $G_X(x)$, etc. The functions with the lowercase letters are *probability density functions* and those with uppercase letters are *cumulative distribution functions.*

For **discrete random variables**, the probability density function has the properties

1. $f_X(x_k) = P(X = x_k)$

2. $\sum_{k=1}^{\infty} f_X(x_k) = 1$

And the cumulative distribution function has the properties

1. $F_X(x_k) = P(X \le x_k)$

2. $F_X(x) = \sum_{x_k \le x} f_X(x_k)$

As an example consider the probability distribution functions $f_X(x)$ and $F_X(x)$ of the sum of the numbers when a pair of dice is tossed.

Experiment: Toss two identical dice.

Sample space:

$$S = \left\{ \begin{matrix} 1,1 & 1,2 & 1,3 & 1,4 & 1,5 & 1,6 \\ 2,1 & 2,2 & 2,3 & 2,4 & 2,5 & 2,6 \\ 3,1 & 3,2 & 3,3 & 3,4 & 3,5 & 3,6 \\ 4,1 & 4,2 & 4,3 & 4,4 & 4,5 & 4,6 \\ 5,1 & 5,2 & 5,3 & 5,4 & 5,5 & 5,6 \\ 6,1 & 6,2 & 6,3 & 6,4 & 6,5 & 6,6 \end{matrix} \right\}$$

Random Variable: X, the total of the two numbers

The probability the random variable X assumes the possible values $x = 2, 3, 4, \dots, 12$ is given in Table A1

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Table A1. Table of probabilities

Note that the values of x exhaust all possible cases and hence the probabilities add to 1

The probability density function $f_X(x)$ can be deduced from Table A1

$$f_X(x) = \frac{6 - |x - 7|}{36}, \qquad x = 2, 3, 4, \dots, 12$$

Probability distributions are often shown in graphical form. For discrete random variables, probability distributions are generally shown in the form of histograms consisting of series of rectangles associated with values of the random variable. The width of each rectangle is one unit and the height is the probability given by the function $f_X(x)$ and the sum of the areas of all the rectangles is 1.

Figure A1 shows the Probability histogram for the random variable X, the sum of the numbers when a pair of dice is tossed.

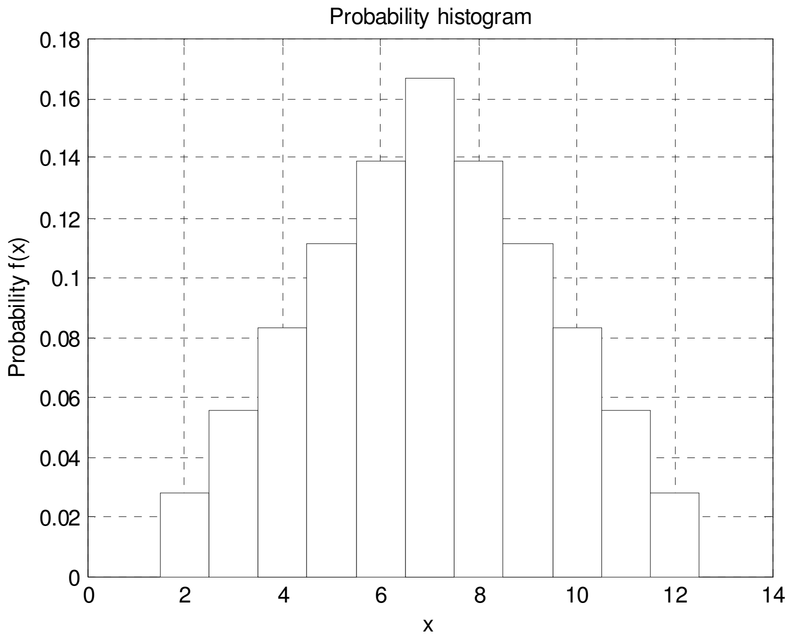

Figure A1 Probability histogram

Figure A2 shows the cumulative distribution function $F_X\left(x\right) = \sum_{x_k \leq x} f_X\left(x_k\right)$ for the random variable X, the sum of the numbers when a pair of dice is tossed.

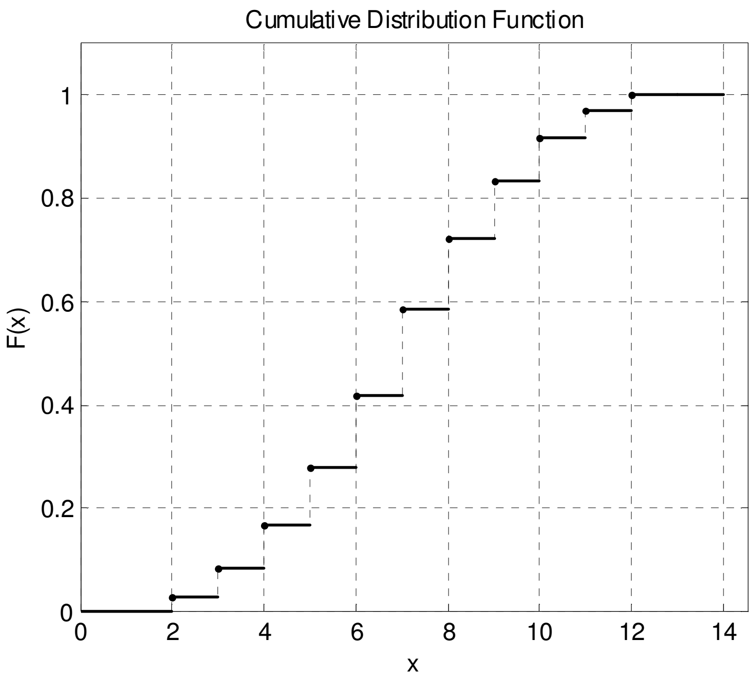

Figure A2. Cumulative distribution function. [The dots at the left ends of the line segments indicate the value of $F_X\left(x\right)$ at those values of x.

For **continuous random variables**, the probability distribution functions $f_X(x)$ and $F_X(x)$ are curves, which may take various forms depending on the nature of the random variable. Probability density functions $f_X(x)$ that are used in practice to model the behaviour of continuous random variables are always positive and the total area under its curve, bounded by the x-axis, is equal to one. These density functions have the following properties

1. $f_X(x) \geq 0$ for any value of x
2. $\int_{-\infty}^{+\infty} f_X(x)\,dx = 1$

The probability that a random variable X lies between any two values $x = a$ and $x = b$ is the area under the density curve between those two values and is found by methods of integral calculus

$$P(a < X < b) = \int_a^b f_X(x)\,dx \tag{78}$$

The equations of the density functions $f_X(x)$ are usually complicated and areas under their curves are found from tables. In many scientific studies, the Normal probability density function is the usual model for the behaviour of measurements (regarded as random variables) and the probability density function is (Kreyszig, 1970, p. 107)

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{79}$$

μ and σ are the mean and standard deviation respectively of the infinite population of x and Figure A3 shows a plot of the Normal probability density curve for $\mu = 2.0$ and $\sigma = 2.5$.

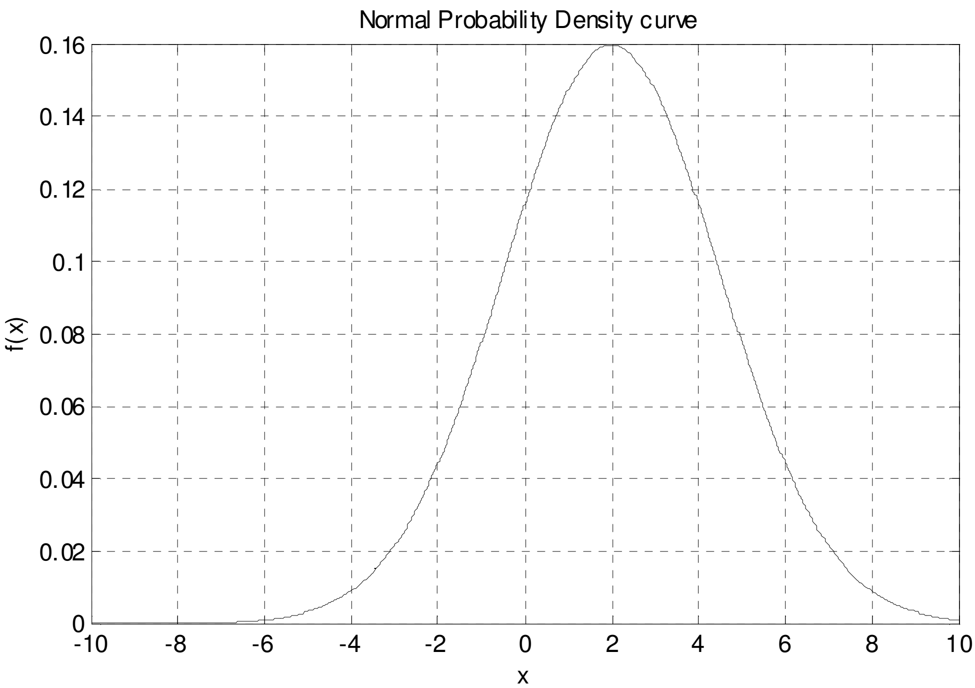

Figure A3. Normal probability density function for $\mu = 2.0$ and $\sigma = 2.5$

For continuous random variables *X,* the cumulative distribution function $F_X(x)$ has the following properties

1. $F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(x)\,dx$

2. $\frac{d}{dx} F_X(x) = f_X(x)$

In many scientific studies, the Normal distribution is the usual model for the behaviour of measurements and the cumulative distribution function is (Kreyszig, 1970, p. 108)

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \tag{80}$$

The probability that X assumes any value in an interval $a < X < b$ is

$$P(a < X < b) = F_X(b) - F_X(a) = \frac{1}{\sigma\sqrt{2\pi}} \int_{a}^{b} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \tag{81}$$

Figure A4 shows a plot of the Normal cumulative distribution curve for $\mu = 2.0$ and $\sigma = 2.5$.

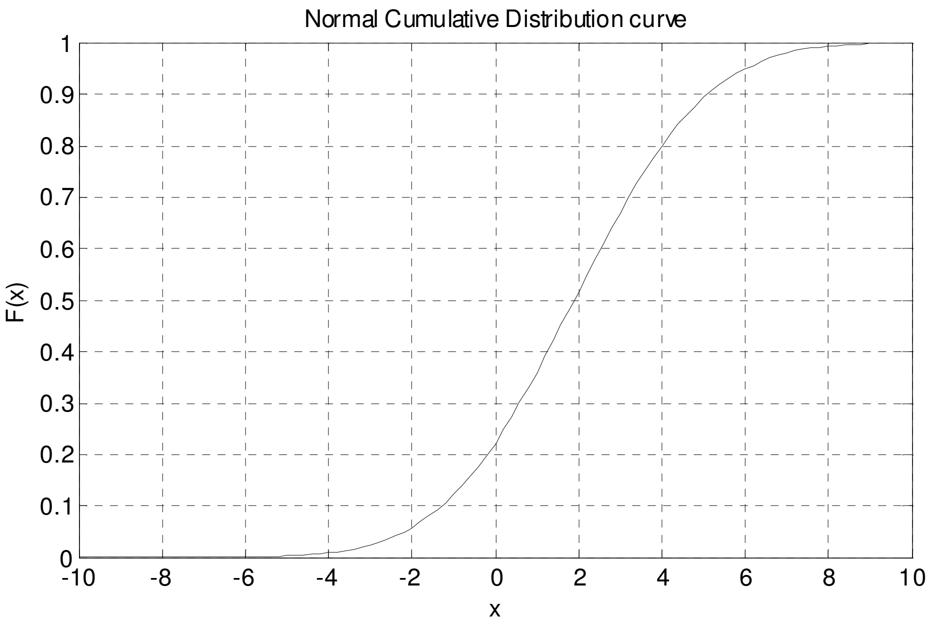

Figure A4. Normal cumulative distribution function for $\mu = 2.0$ and $\sigma = 2.5$

Expectations

The *expectation* $E\{X\}$ of a random variable X is defined as the average value μ_X of the variable over all possible values. It is computed by taking the sum of all possible values of $X = x$ multiplied by its corresponding probability. In the case of a *discrete* random variable the expectation is given by

$$E\{X\} = \mu_X = \sum_{k=1}^{N} x_k P(x_k) \tag{82}$$

Equation (82) is a general expression from which we can obtain the usual expression for the arithmetic mean

$$\mu = \frac{1}{N} \sum_{k=1}^{N} x_k \tag{83}$$

If there are N possible values x_k of the random variable X, each having equal probability $P\left(x_k\right) = 1/N$ (which is a constant), then the expectation computed from (82) is identical to the arithmetic mean of the N values of x_k from (83).

In the case of a *continuous* random variable the expectation is given by

$$E\left\{X\right\} = \mu_X = \int_{-\infty}^{+\infty} x\, f_X\left(x\right) dx \tag{84}$$

This relationship may be extended to a more general form if we consider the expectation of a function $g\left(X\right)$ of a random variable X whose probability density function is $f_X\left(x\right)$. In this case

$$E\left\{g\left(X\right)\right\} = \int_{-\infty}^{+\infty} g\left(x\right) f_X\left(x\right) dx \tag{85}$$

Extending (85) to the case of two random variables X and Y

$$E\left\{g\left(X,Y\right)\right\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g\left(x,y\right) f_{XY}\left(x,y\right) dx\, dy$$

Similarly for n random variables

$$E\left\{g\left(X_1,\ X_2,\ldots,X_n\right)\right\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} g\left(x_1,x_2,\ldots,x_n\right) f_X\left(x_1,x_2,\ldots,x_n\right) dx_1\, dx_2 \ldots dx_n \tag{86}$$

Expressing (86) in matrix notation gives a general form of the expected value of a multivariate function $g\left(\mathbf{X}\right)$ as

$$E\left\{g\left(\mathbf{X}\right)\right\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} g\left(\mathbf{x}\right) f_{\mathbf{X}}\left(\mathbf{x}\right) d\mathbf{x} \tag{87}$$

where $f_{\mathbf{X}}\left(\mathbf{x}\right)$ is the multivariate probability density function.

There are some rules that are useful in calculating expectations. They are given here without proof but can be found in many statistical texts. With a and b as constants and X and Y as random variables

$$E\left\{a\right\} = a$$

$$E\left\{aX\right\} = a\,E\left\{X\right\}$$

$$E\left\{aX + b\right\} = a\,E\left\{X\right\} + b$$

$$E\left\{g\left(X\right) \pm h\left(X\right)\right\} = E\left\{g\left(X\right)\right\} \pm E\left\{h\left(X\right)\right\}$$

$$E\left\{g\left(X,Y\right) \pm h\left(X,Y\right)\right\} = E\left\{g\left(X,Y\right)\right\} \pm E\left\{h\left(X,Y\right)\right\}$$

Special Mathematical Expectations

The *mean* of a random variable

$$\mu_X = E\{X\} = \int_{-\infty}^{+\infty} x\, f_X(x)\, dx \tag{88}$$

The *mean vector* $\mathbf{m}_X$ of a multivariate distribution is

$$\mathbf{m}_X = \begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \\ \mu_{X_3} \\ \vdots \end{bmatrix} = \begin{bmatrix} E(X_1) \\ E(X_2) \\ E(X_3) \\ \vdots \end{bmatrix} = E \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \end{bmatrix} = E\{\mathbf{X}\} \tag{89}$$

$\mathbf{m}_X$ can be taken as representing the mean of a multivariate probability density function.

The *variance* of a random variable

$$\sigma_X^2 = E\left\{(X - \mu_X)^2\right\} = \int_{-\infty}^{+\infty} (x - \mu_x)^2 f_X(x)\, dx \tag{90}$$

The *covariance* between two random variables X and Y is

$$\sigma_{XY} = E\{(X - \mu_X)(Y - \mu_Y)\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x - \mu_x)(y - \mu_y) f_{XY}(x,y)\, dx\, dy \tag{91}$$

Equation (91) can be expanded to give

$$\begin{aligned} \sigma_{XY} &= E\{(X - \mu_X)(Y - \mu_Y)\} \\ &= E\{XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y\} \\ &= E\{XY\} - E\{X\mu_Y\} - E\{Y\mu_X\} + E\{\mu_X\mu_Y\} \\ &= E\{XY\} - \mu_Y E\{X\} - \mu_X E\{Y\} + \mu_X\mu_Y \\ &= E\{XY\} - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y \\ &= E\{XY\} - \mu_X\mu_Y \end{aligned}$$

If the random variables X and Y are independent, the expectation of the product is equal to the product of the expectations, i.e., $E\{XY\} = E\{X\}E\{Y\}$. Since the expected values of X and Y are the means μ_X and μ_Y then $E\{XY\} = \mu_X\mu_Y$ if X and Y are independent. Substituting this result into the expansion above shows that the covariance σ_{XY} is zero if X and Y are independent.

For a multivariate function, variances and covariances of the random variables $\mathbf{X}$ is given by the matrix equation

$$\mathbf{V}_{XX} = E\left\{[\mathbf{X} - \mathbf{m}_X][\mathbf{X} - \mathbf{m}_Y]^T\right\} \tag{92}$$

$\mathbf{V}_{XX}$ is a symmetric matrix known as the variance-covariance matrix and its general form can be seen when (92) is expanded

$$\mathbf{V}_{XX} = E\left\{ \begin{bmatrix} X_1 - \mu_{X_1} \\ X_2 - \mu_{X_2} \\ \vdots \\ X_n - \mu_{X_n} \end{bmatrix} \begin{bmatrix} X_1 - \mu_{X_1} & X_2 - \mu_{X_2} & \cdots & X_n - \mu_{X_n} \end{bmatrix} \right\}$$

giving
$$\mathbf{V}_{XX} = \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_n} \\ \sigma_{X_2X_1} & \sigma_{X_1}^2 & \cdots & \sigma_{X_2X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_nX_1} & \sigma_{X_nX_2} & \cdots & \sigma_{X_1}^2 \end{bmatrix} \tag{93}$$

Law of Propagation of Variances for Linear Functions

Consider two vectors of random variables $\mathbf{x} = \begin{bmatrix} X_1 & X_2 & \cdots & X_n \end{bmatrix}^T$ and $\mathbf{y} = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_n \end{bmatrix}^T$ that are linearly related by the matrix equation

$$\mathbf{y} = \mathbf{Ax} + \mathbf{b} \tag{94}$$

where $\mathbf{A}$ is a coefficient matrix and $\mathbf{b}$ is a vector of constants. Then, using the rules for expectations developed above we may write an expression for the mean $\mathbf{m}_Y$ using (89)

$$\begin{aligned} \mathbf{m}_Y &= E\left\{\mathbf{y}\right\} \\ &= E\left\{\mathbf{Ax} + \mathbf{b}\right\} \\ &= E\left\{\mathbf{Ax}\right\} + E\left\{\mathbf{b}\right\} \\ &= \mathbf{A}\,E\left\{\mathbf{x}\right\} + \mathbf{b} \\ &= \mathbf{Am}_X + \mathbf{b} \end{aligned}$$

Using (92), the variance-covariance matrix $\mathbf{V}_{yy}$ is given by

$$\begin{aligned} \mathbf{V}_{yy} &= E\left\{\left(\mathbf{y} - \mathbf{m}_y\right)\left(\mathbf{y} - \mathbf{m}_y\right)^T\right\} \\ &= E\left\{\left(\mathbf{Ax} + \mathbf{b} - \mathbf{Am}_x - \mathbf{b}\right)\left(\mathbf{Ax} + \mathbf{b} - \mathbf{Am}_x - \mathbf{b}\right)^T\right\} \\ &= E\left\{\left(\mathbf{Ax} - \mathbf{Am}_x\right)\left(\mathbf{Ax} - \mathbf{Am}_x\right)^T\right\} \\ &= E\left\{\mathbf{A}\left(\mathbf{x} - \mathbf{m}_x\right)\left(\mathbf{A}\left(\mathbf{x} - \mathbf{m}_x\right)\right)^T\right\} \\ &= E\left\{\mathbf{A}\left(\mathbf{x} - \mathbf{m}_x\right)\left(\mathbf{x} - \mathbf{m}_x\right)^T \mathbf{A}^T\right\} \\ &= \mathbf{A}\,E\left\{\left(\mathbf{x} - \mathbf{m}_x\right)\left(\mathbf{x} - \mathbf{m}_x\right)^T\right\}\mathbf{A}^T \\ &= \mathbf{AV}_{xx}\mathbf{A}^T \end{aligned}$$

or

If $\mathbf{y} = \mathbf{Ax} + \mathbf{b}$ and $\mathbf{y}$ and $\mathbf{x}$ are random variables linearly related then

$$\mathbf{V}_{yy} = \mathbf{AV}_{xx}\mathbf{A}^T \tag{95}$$

Equation (95) is known as the *Law of Propagation of Variances.*

Law of Propagation of Variances for Non-Linear Functions

In many practical applications of variance propagation the random variables in $\mathbf{x}$ and $\mathbf{y}$ are nonlinearly related, i.e.,

$$\mathbf{y} = f(\mathbf{x}) \tag{96}$$

In such cases, we can expand the function on the right-hand-side of (96) using Taylor's theorem.

For a non-linear function of a <u>single</u> variable Taylor's theorem may be expressed in the following form

$$\begin{aligned} f(x) = f(a) &+ \left.\frac{df}{dx}\right|_a (x-a) + \left.\frac{d^2 f}{dx^2}\right|_a \frac{(x-a)^2}{2!} + \left.\frac{d^3 f}{dx^3}\right|_a \frac{(x-a)^3}{3!} + \cdots \\ &+ \left.\frac{d^{n-1} f}{dx^{n-1}}\right|_a \frac{(x-a)^{n-1}}{(n-1)!} + R_n \end{aligned} \tag{97}$$

where R_n is the remainder after n terms and $\lim_{n\to\infty} R_n = 0$ for $f(x)$ about $x = a$ and $\left.\frac{df}{dx}\right|_a$, $\left.\frac{d^2 f}{dx^2}\right|_a$ etc. are derivatives of the function $f(x)$ evaluated at $x = a$.

For a non-linear function of <u>two</u> random variables, say $\phi = f(x, y)$, the Taylor series expansion of the function ϕ about $x = a$ and $y = b$ is

$$\begin{aligned} \phi = f(a,b) &+ \left.\frac{\partial f}{\partial x}\right|_{a,b} (x-a) + \left.\frac{\partial f}{\partial y}\right|_{a,b} (y-b) \\ &+ \frac{1}{2!}\left\{ \left.\frac{\partial^2 f}{\partial x^2}\right|_{a,b} (x-a)^2 + \left.\frac{\partial^2 f}{\partial y^2}\right|_{a,b} (y-b)^2 + \left.\frac{\partial f}{\partial x}\right|_{a,b} \left.\frac{\partial f}{\partial y}\right|_{a,b} (x-a)(y-b) \right\} + \cdots \end{aligned} \tag{98}$$

where $f(a,b)$ is the function ϕ evaluated at $x = a$ and $y = b$, and $\left.\frac{\partial f}{\partial x}\right|_{a,b}$, $\left.\frac{\partial f}{\partial y}\right|_{a,b}$, $\left.\frac{\partial^2 f}{\partial x^2}\right|_{a,b}$ etc are partial derivatives of the function ϕ evaluated at $x = a$ and $y = b$.

Extending to n random variables, we may write a Taylor series approximation of the function $f(\mathbf{x})$ as a matrix equation

$$f(\mathbf{x}) = f(\mathbf{x}^0) + \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}^0} (\mathbf{x} - \mathbf{x}^0) + \text{ higher order terms} \tag{99}$$

where $f(\mathbf{x}^0)$ is the function evaluated at the approximate values $\mathbf{x}^0$ and $\left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}^0}$ are the partial derivatives evaluated at approximations $\mathbf{x}^0$.

Replacing $f(\mathbf{x})$ in (96) by its Taylor series approximation, ignoring higher order terms, gives

$$\mathbf{y} = f(\mathbf{x}) = f(\mathbf{x}^0) + \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}^0} (\mathbf{x} - \mathbf{x}^0) \tag{100}$$

Then, using the rules for expectations

$$
\begin{aligned}
\mathbf{m}_y &= E\left\{\mathbf{y}\right\} \\
&= E\left\{ f\left(\mathbf{x}^0\right) + \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}^0} \left(\mathbf{x} - \mathbf{x}^0\right) \right\} \\
&= E\left\{ f\left(\mathbf{x}^0\right) \right\} + E\left\{ \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}^0} \left(\mathbf{x} - \mathbf{x}^0\right) \right\} \\
&= f\left(\mathbf{x}^0\right) + \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}^0} E\left\{ \left(\mathbf{x} - \mathbf{x}^0\right) \right\} \\
&= f\left(\mathbf{x}^0\right) + \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}^0} \left(E\left\{\mathbf{x}\right\} - E\left\{\mathbf{x}^0\right\} \right) \\
&= f\left(\mathbf{x}^0\right) + \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}^0} \left(\mathbf{m}_x - \mathbf{x}^0\right)
\end{aligned}
$$

And

$$
\begin{aligned}
\mathbf{y} - \mathbf{m}_y &= \left[f\left(\mathbf{x}^0\right) + \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}^0} \left(\mathbf{x} - \mathbf{x}^0\right) \right] - \left[f\left(\mathbf{x}^0\right) + \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}^0} \left(\mathbf{m}_x - \mathbf{x}^0\right) \right] \\
&= \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}^0} \left(\mathbf{x} - \mathbf{m}_x\right) \\
&= \mathbf{J}_{yx} \left(\mathbf{x} - \mathbf{m}_x\right)
\end{aligned}
\tag{101}
$$

$\mathbf{J}_{yx}$ is the (m,n) Jacobian matrix of partial derivatives, noting that $\mathbf{y}$ and $\mathbf{x}$ are $(m,1)$ and $(n,1)$ vectors respectively

$$
\mathbf{J}_{yx} = \begin{bmatrix}
\partial y_1 / \partial x_1 & \partial y_1 / \partial x_2 & \cdots & \partial y_1 / \partial x_n \\
\partial y_2 / \partial x_1 & \partial y_2 / \partial x_2 & \cdots & \partial y_2 / \partial x_n \\
\vdots & & & \vdots \\
\partial y_m / \partial x_1 & \partial y_m / \partial x_2 & \cdots & \partial y_m / \partial x_n
\end{bmatrix}
\tag{102}
$$

Using (92), the variance-covariance matrix $\mathbf{V}_{yy}$ is given by

$$
\begin{aligned}
\mathbf{V}_{yy} &= E\left\{ \left(\mathbf{y} - \mathbf{m}_y\right)\left(\mathbf{y} - \mathbf{m}_y\right)^T \right\} \\
&= E\left\{ \left(\mathbf{J}_{yx}\left(\mathbf{x} - \mathbf{m}_x\right)\right)\left(\mathbf{J}_{yx}\left(\mathbf{x} - \mathbf{m}_x\right)\right)^T \right\} \\
&= E\left\{ \mathbf{J}_{yx}\left(\mathbf{x} - \mathbf{m}_x\right)\left(\mathbf{x} - \mathbf{m}_x\right)^T \mathbf{J}_{yx}^T \right\} \\
&= \mathbf{J}_{yx}\, E\left\{ \left(\mathbf{x} - \mathbf{m}_x\right)\left(\mathbf{x} - \mathbf{m}_x\right)^T \right\} \mathbf{J}_{yx}^T \\
&= \mathbf{J}_{yx}\mathbf{V}_{xx}\mathbf{J}_{yx}^T
\end{aligned}
$$

Thus, in a similar manner to above, we may express the Law of Propagation of Variances for non-linear functions of random variables as

If $\mathbf{y} = f\left(\mathbf{x}\right)$ and $\mathbf{y}$ and $\mathbf{x}$ are random variables non-linearly related then

$$
\mathbf{V}_{yy} = \mathbf{J}_{yx}\mathbf{V}_{xx}\mathbf{J}_{yx}^T
\tag{103}
$$

The Special Law of Propagation of Variances

The Law of Propagation of Variances is often expressed as an algebraic equation. For example, if z is a function of two random variables x and y, i.e., $z = f(x,y)$ then the variance of z is

$$\sigma_z^2 = \left(\frac{\partial z}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial z}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial z}{\partial x}\frac{\partial z}{\partial y}\sigma_{xy} \tag{104}$$

Equation (104) can be derived from the general matrix equation (103) in the following manner. Let $z = f(x,y)$ be written as $\mathbf{y} = f(\mathbf{x})$ where $\mathbf{y} = [z]$, a (1,1) matrix and $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ is a (2,1) vector. The variance-covariance matrix of the random vector $\mathbf{x}$ is $\mathbf{V}_{xx} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$, the Jacobian $\mathbf{J}_{yx} = \begin{bmatrix} \dfrac{\partial z}{\partial x} & \dfrac{\partial z}{\partial y} \end{bmatrix}$ and the variance-covariance matrix $\mathbf{V}_{yy}$ which contains the single element σ_z^2 is given by

$$\mathbf{V}_{yy} = \left[\sigma_z^2\right] = \begin{bmatrix} \dfrac{\partial z}{\partial x} & \dfrac{\partial z}{\partial y} \end{bmatrix} \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \begin{bmatrix} \dfrac{\partial z}{\partial x} \\ \dfrac{\partial z}{\partial y} \end{bmatrix}$$

Expanding this equation gives (104).

In the case where the random variables in $\mathbf{x}$ are independent, i.e., their covariances are zero; we have the Special Law of Propagation of Variances. For the case of $z = f(x,y)$ where the random variables x and y are independent, the *Special Law of Propagation of Variances* is written as

If $z = f(x,y)$ and x and y are independent random variables then

$$\sigma_z^2 = \left(\frac{\partial z}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial z}{\partial y}\right)^2 \sigma_y^2 \tag{105}$$

APPENDIX B: MATLAB function *logistic.m*

http://www.cs.cmu.edu/~ggordon/IRLS-example/logistic.m

```
% function x = logistic(a, y, w, ridge, param)
%
% Logistic regression.  Design matrix A, targets Y, optional instance
% weights W, optional ridge term RIDGE, optional parameters object PARAM.
%
% W is a vector with length equal to the number of training examples; RIDGE
% can be either a vector with length equal to the number of regressors, or
% a scalar (the latter being synonymous to a vector with all entries the
% same).
%
% PARAM has fields PARAM.MAXITER (an iteration limit), PARAM.VERBOSE
% (whether to print diagnostic information), PARAM.EPSILON (used to test
% convergence), and PARAM.MAXPRINT (how many regression coefficients to
% print if VERBOSE==1).
%
% Model is
%
%   E(Y) = 1 ./ (1+exp(-A*X))
%
% Outputs are regression coefficients X.
%
% Copyright 2007 Geoffrey J. Gordon
%
% This program is free software: you can redistribute it and/or modify
% it under the terms of the GNU General Public License as published by
% the Free Software Foundation, either version 3 of the License, or (at
% your option) any later version.
%
% This program is distributed in the hope that it will be useful, but
% WITHOUT ANY WARRANTY; without even the implied warranty of
% MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the GNU
% General Public License for more details.
%
% You should have received a copy of the GNU General Public License
% along with this program.  If not, see <http://www.gnu.org/licenses/>.

function x = logistic(a, y, w, ridge, param)

% process parameters

[n, m] = size(a);

if ((nargin < 3) || (isempty(w)))
  w = ones(n, 1);
end

if ((nargin < 4) || (isempty(ridge)))
  ridge = 1e-5;
end

if (nargin < 5)
  param = [];
end

if (length(ridge) == 1)
    ridgemat = speye(m) * ridge;
elseif (length(ridge(:)) == m)
    ridgemat = spdiags(ridge(:), 0, m, m);
else
    error('ridge weight vector should be length 1 or %d', m);
```

```
end

if (~isfield(param, 'maxiter'))
  param.maxiter = 200;
end

if (~isfield(param, 'verbose'))
  param.verbose = 0;
end

if (~isfield(param, 'epsilon'))
  param.epsilon = 1e-10;
end

if (~isfield(param, 'maxprint'))
  param.maxprint = 5;
end

% do the regression

x = zeros(m,1);
oldexpy = -ones(size(y));
for iter = 1:param.maxiter

  adjy = a * x;
  expy = 1 ./ (1 + exp(-adjy));
  deriv = expy .* (1-expy);
  wadjy = w .* (deriv .* adjy + (y-expy));
  weights = spdiags(deriv .* w, 0, n, n);

  x = inv(a' * weights * a + ridgemat) * a' * wadjy;

  if (param.verbose)
    len = min(param.maxprint, length(x));
    fprintf('%3d: [',iter);
    fprintf(' %g', x(1:len));
    if (len < length(x))
      fprintf(' ... ');
    end
    fprintf(' ]\n');
  end

  if (sum(abs(expy-oldexpy)) < n*param.epsilon)
    if (param.verbose)
      fprintf('Converged.\n');
    end
    return;
  end

  oldexpy = expy;

end

warning('logistic:notconverged', 'Failed to converge');
```

Useage Example

http://www.cs.cmu.edu/~ggordon/IRLS-example/logistic-ex.txt

```
>> a = randn(500,5);
>> x = 2*randn(5,1);
>> y = (rand(500,1) < 1./(1+exp(-a*x)));
>> xhat = logistic(a, y, [], [], struct('verbose', 1))
  1: [ -0.842889 -0.959492 0.843404 0.198022 0.199493 ]
  2: [ -1.55055 -1.75901 1.57622 0.360507 0.398254 ]
  3: [ -2.35678 -2.71685 2.4373 0.545032 0.62605 ]
  4: [ -3.20879 -3.74533 3.35828 0.740927 0.854976 ]
  5: [ -3.86696 -4.54162 4.0753 0.899695 1.02575 ]
  6: [ -4.12265 -4.85136 4.35625 0.964904 1.09126 ]
  7: [ -4.14969 -4.88416 4.38617 0.972078 1.09818 ]
  8: [ -4.14995 -4.88447 4.38646 0.972149 1.09825 ]
  9: [ -4.14995 -4.88447 4.38646 0.972149 1.09825 ]
 10: [ -4.14995 -4.88447 4.38646 0.972149 1.09825 ]
 11: [ -4.14995 -4.88447 4.38646 0.972149 1.09825 ]
Converged.

xhat =

   -4.1499
   -4.8845
    4.3865
    0.9721
    1.0982

>> x

x =

   -3.9412
   -4.0619
    3.6705
    1.1123
    0.9645
```