McGill University

Department of Epidemiology
and Biostatistics

EPIB – 621

Data Analysis in the Health Sciences

Lawrence Joseph

Data Analysis for the Health Sciences – EPIB-621 – 4 credits

Instructors:	Lawrence Joseph
Email address:	Lawrence.Joseph@mcgill.ca
Course Home page:	www.medicine.mcgill.ca/epidemiology/Joseph/EPIB-621.html
Telephone:	934-1934 X 44713
Address:	McGill University Health Centre 2155 Guy Street, #343 Montreal, Quebec, Canada, H3H 2R9

Brief Course Outline: Univariate and multivariate statistical techniques for continuous and dichotomous outcomes. Focus is on multiple linear and logistic regression models. Additional topics will include adjusting for missing data, measurement error and hierarchical (random effects) models. All material will be taught from both Bayesian and frequentist viewpoints. R and WinBUGS software will be used throughout the course.

Place and Time: January 7 to April 10, 2019. Mondays and Wednesdays, 11:30–1:30 PM. All lectures and exams take place in Room 1034, McIntyre Medical Building.

Assessment: Assignments $5 \times 4\%$ each $= 20\%$, Midterm Exam $= 30\%$, Final Exam $= 50\%$.

Prerequisites: Previous courses in differential and integral calculus, and EPIB 607 or equivalent.

Midterm Exam: February 20, 2019, 11:30 PM to 1:30 PM, Room 1034, McIntyre.
Final Exam: April 10, 2019, 11:30 PM to 1:30 PM, Room 1034, McIntyre.

Please note that both exams are open book.

Equipment:

- Access to a computer with R and WinBUGS software loaded on (both are freeware) is required. Computers in the basement of Purvis Hall can be used, as needed.
- A hand calculator with logarithm and exponential functions will be required for exams. Use of computers or other similar devices will not be allowed during exams.

Reference material: No single textbook will be exactly followed, but most material covered in the course is included in the following textbooks. The lectures will follow the course pack (see link below).

- Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li. Applied Linear Statistical Models, 5th Edition, McGraw-Hill, 2005.
- David W. Hosmer and Stanley Lemeshow, Applied Logistic Regression, 2nd Edition Wiley, 2000.
- A. Gelman, J. Carlin, H. Stern and D. Rubin, Bayesian Data Analysis, 2nd Edition, Chapman and Hall, 2003.
- Woodworth G, Biostatistics: A Bayesian Introduction, Wiley, 2004.
- Course pack is downloadable from:

  http://www.medicine.mcgill.ca/epidemiology/Joseph/

  Follow the link to the 621 course, then click on "Course outline". Each lecture is available in pdf form by clicking on the lecture title. Supplementary material is available for some lectures, listed separately.

  We strongly suggest that you print out a copy of the lectures, bringing them to class, perhaps collecting the pages into a binder. This is the material I will actually follow, on a day-to-day basis, some of the print may be difficult to view on the screen, and you will almost surely want to take your own notes on top of some of the pages. The text books are for reference, although you will likely wish to buy a copy of one or more of these, not only for this course, but as excellent references for all your future analysis work.

Overview of Content: The course will be divided into four main sections:

1. **Review of basic univariate statistics:** Inferences for means and proportions, simple linear regression, p-values and confidence intervals. Includes introduction to the Bayesian viewpoint for these three basic types of analyses, including Bayesian posterior distributions and credible intervals. The R software package will be introduced.

2. **Linear regression:** Linear regression for two or more explanatory (X) variables, including polynomial terms, use of dummy variables, inference for regression parameters from frequentist and Bayesian viewpoints, residuals, addressing confounding and the use of interaction terms, model selection, goodness of fit, predictions using regression equations, hierarchical (random effects) models, programming in R and WinBUGS.

3. **Logistic regression:** Logistic regression for one or more explanatory variables, including use of dummy variables, inference for logistic regression parameters and odds ratios from frequentist and Bayesian viewpoints, addressing confounding and the use of interaction terms, model selection, goodness of fit, the inverse logit for predictions using logistic regression equations, hierarchical (random effects) models, programming in R and WinBUGS.

4. **Additional topics (as time permits):** Adjusting for missing data and measurement error.

See detailed outline on page 4 for the topics to be covered in each lecture.

How to get the most from this course: We will be covering a lot of material, much of which will be totally new, and it will most likely all appear to go by very fast. **Please never hesitate** to ask questions in class and contact us via email with questions outside of class. You will get the most out of this course if you keep up with the material on a day-to-day basis, and ask questions about what you do not understand right away, rather than waiting until later, when it may become impossible to catch up. This applies both to the "theoretical" material and to the software packages we will use.

Data Analysis in the Health Sciences

Course Outline – EPIB–621

Date	**Topic Covered**
Mon Jan 7	Introduction/Motivation/Evaluation/Scope/Background
Wed Jan 9	Frequentist inferences for means and proportions
Mon Jan 14	Bayesian inferences for means and proportions
Wed Jan 16	Introduction to R
Mon Jan 21	Simple linear regression: one variable
Wed Jan 23	Linear regression with two or more variables: basic concepts
Mon Jan 28	Dummy variables in linear regression
Wed Jan 30	Confounding and collinearity in linear regression
Mon Feb 4	Interaction terms and prediction in linear regression
Wed Feb 6	Goodness of fit in linear regression
Mon Feb 11	Bayesian inference for linear regression models
Wed Feb 13	Model selection and predictions in linear regression
Mon Feb 18	Review
Wed Feb 20	Midterm Exam
Mon Feb 25	Hierarchical/random effects linear models
Wed Feb 27	Introduction to logistic regression: Univariate
Mon Mar 4	No Class – Spring Break
Wed Mar 8	No Class – Spring Break
Mon Mar 11	Introduction to logistic regression: Multivariate
Wed Mar 13	Confounding and collinearity in logistic regression
Mon Mar 18	Goodness of fit in logistic regression
Wed Mar 20	Bayesian analysis of logistic regression models
Mon Mar 25	Model selection in logistic regression
Wed Mar 27	Hierarchical/random effects logistic regression
Mon April 1	Missing data
Wed Apr 3	Measurement error
Mon Apr 8	Review
Wed Apr 10	Final Exam

Mathematical Background:

Functions including exp, log, logit, inverse logit, lines, derivatives, integrals

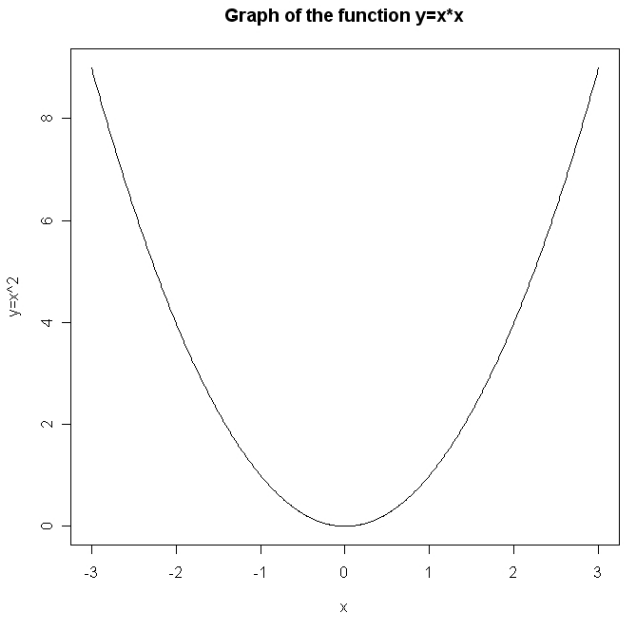
Note: The following are *very* non-rigorous definitions designed to suit the purpose of our course. Refer to any calculus textbook for the exact definitions and/or more information.

Functions: For our purposes, a *function* assigns a unique numerical value to each number in a specified set. For example, the function

$$f(x) = x^2, \ -\infty < x < +\infty$$

assigns the value x^2 to each x, $-\infty < x < +\infty$. Thus $x = 1$ is assigned the value 1, $x = 2$ is assigned the value 4, and $x = -2.1$ is assigned the value $+4.41$, etc. A function is defined over a set of values, which here is the set of all real numbers.

Functions are often easily understood by looking at the *graph* of the function.

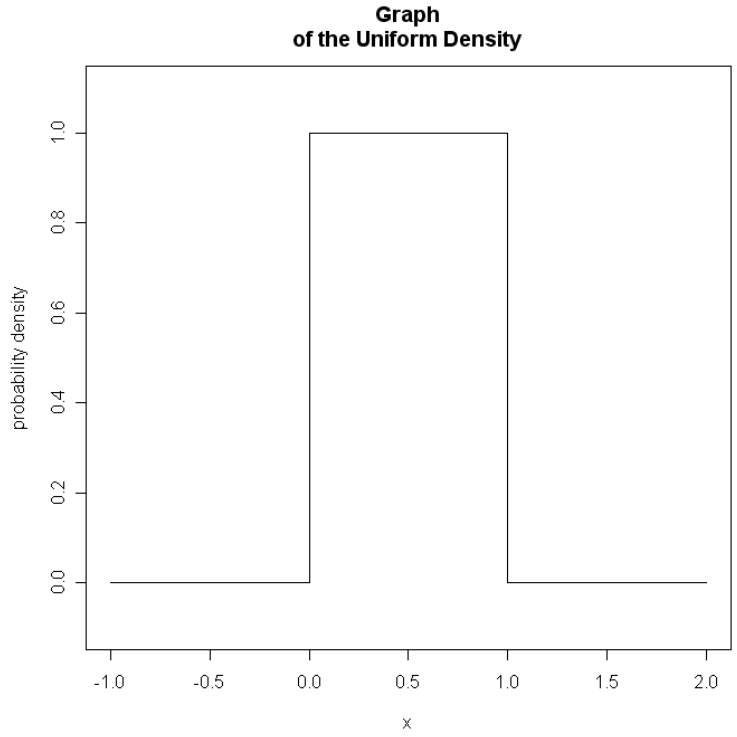


Functions are used throughout statistics to describe probability (density) functions, linear and logistic regression models, and many other places. Let's look at some examples:

(i) The Uniform probability (density) function describes the experiment of choosing a random number between 0 and 1. The function is

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise,} \end{cases}$$

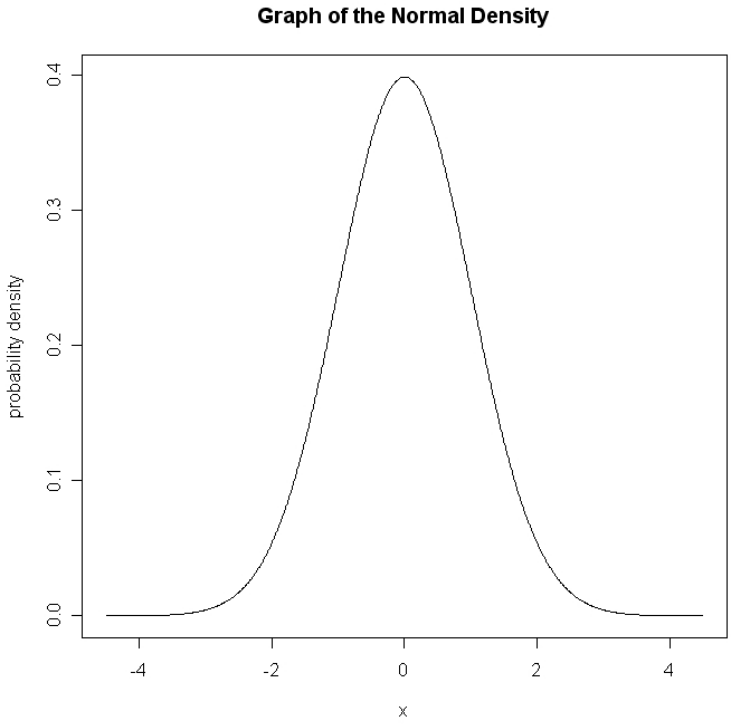
and the graph is shown below:



(ii) The standard Normal probability (density) function is used extensively in virtually every discipline where statistics are used, including medicine. The function is

$$f(x) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{x^2}{2}\right\}, \ -\infty < x < +\infty$$

and the graph is shown below:

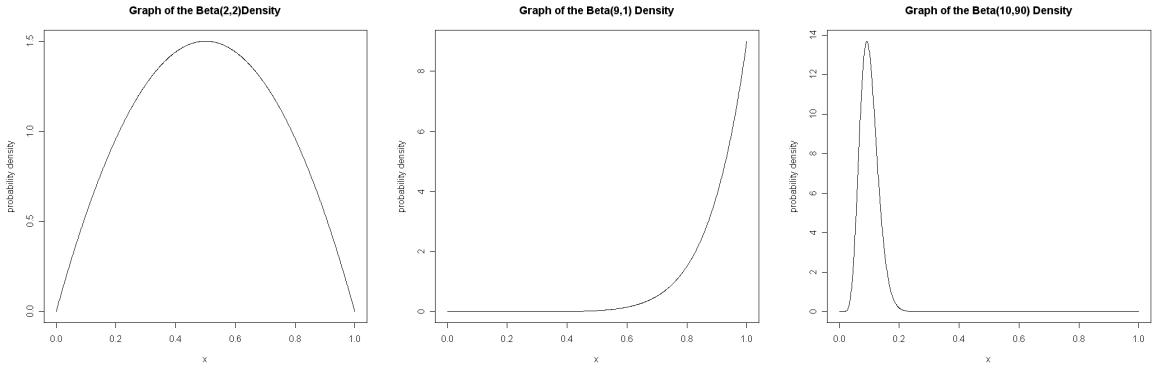


(iii) Another very common density used in Bayesian analysis is the beta. As we will see later in the course, it is typically used in problems involving proportions. Note that its range is between 0 and 1, very convenient for proportions. The function for the beta density is

$$f(\theta) = \begin{cases} \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, & 0 \le \theta \le 1,\ \alpha,\beta > 0, \quad \text{and} \\ 0, & \text{otherwise,} \end{cases} .$$

[$B(\alpha,\beta)$ represents the Beta function evaluated at (α,β). It is simply the normalizing constant that is necessary to make the density integrate to one, that is, $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$.] Some graphs of beta densities are shown below.

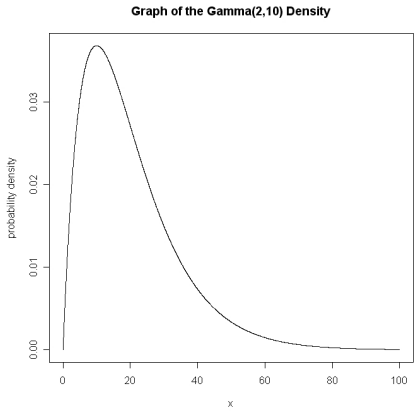
Note the flexibility of this family of distributions.



(iv) Yet another useful distribution is the gamma, which is sometimes used to model normal variances (or, more accurately, as we will see, the inverse of normal variances, known as the precision, i.e., precision = 1/variance). The gamma density is given by

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1}, \ for\ x > 0\ .$$

A typical gamma graph is:

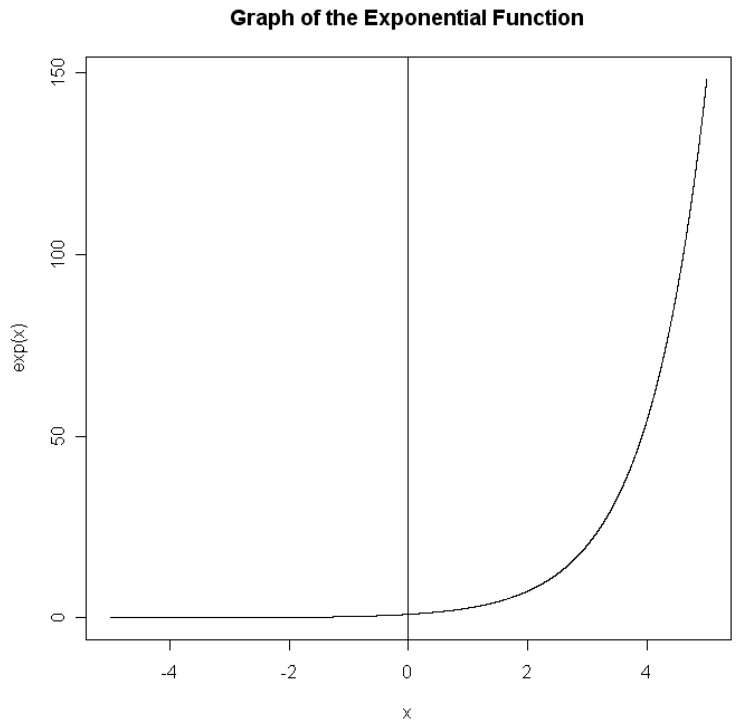


Aside from probability densities, several functions are of particular interest to us in this course.

(v) The exponential function is defined by

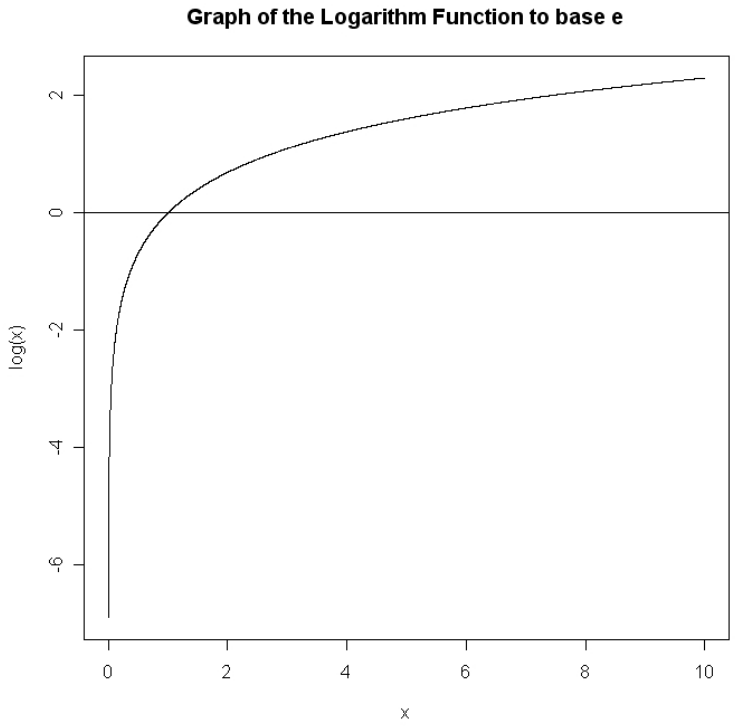
$$f(x) = exp(x) = e^x$$

where e is the constant 2.718282.... It's graph looks like this (note that $\exp(0) = 1$):



We already saw a use of the exp function in the definition of the normal density. It will also be used extensively in logistic regression.

(vi) The log function is the inverse function to the exponential, essentially meaning that $\exp(\log(x)) = x$ and $\log(\exp(x)) = x$ (but watch the range). In general, the logarithm $\log_b(x)$ for a base b and a number x is defined to be the inverse function of taking b to the power x. Here is the graph for base $e = 2.718282$ (note that $\log(1) = 0$):

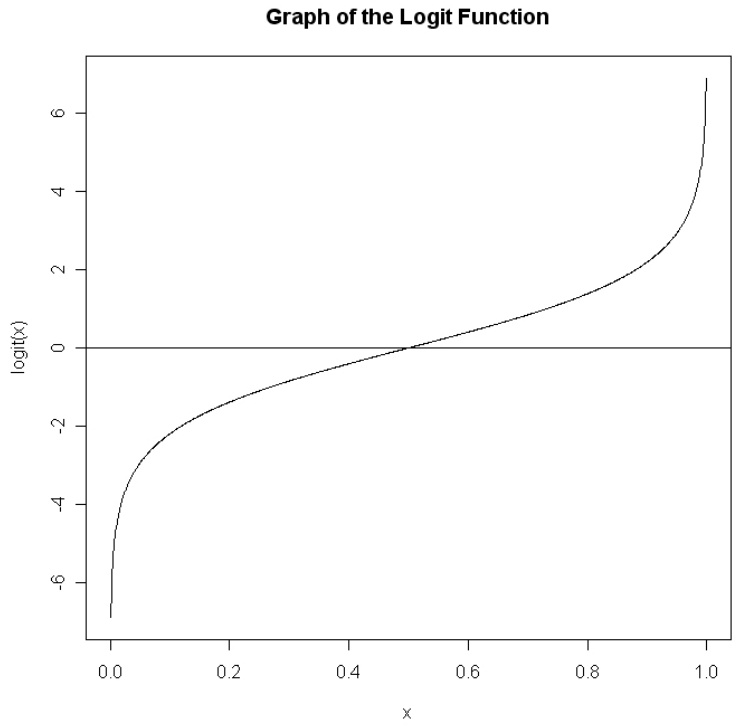


As we will now see, the log function is also used extensively in logistic regression.

(vii) The logit function is defined by

$$f(x) = logit(x) = \log\left(\frac{x}{1-x}\right)$$

It's graph looks like this (note that $logit(0.5) = 0$):

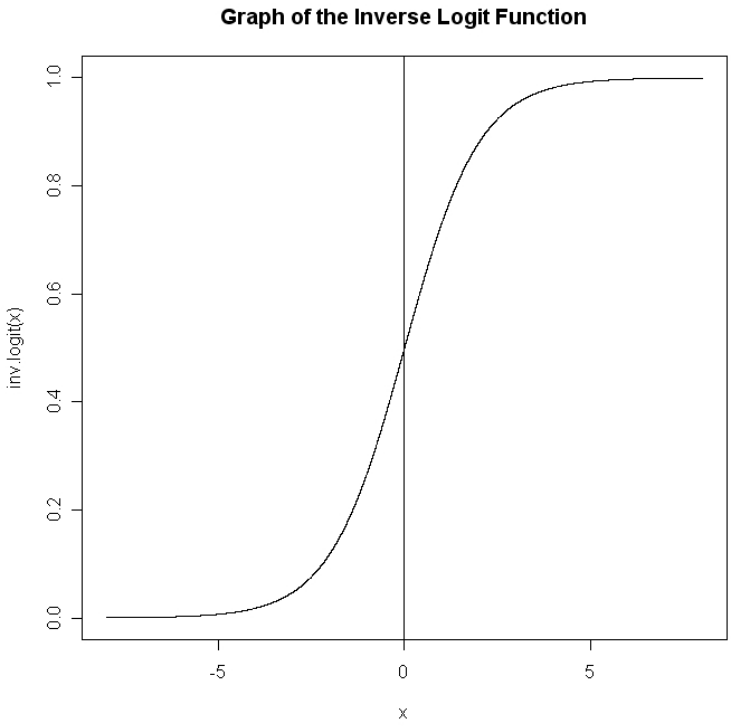


As we will see later in the course, this is the basic function behind logistic regression.

(viii) The inverse logit function is (no surprise here!) the inverse function to the logit.

$$f(x) = inv.logit(x) = \frac{\exp(x)}{1 + \exp(x)}$$

It's graph looks like this (note that $inv.logit(0) = 0.5$):

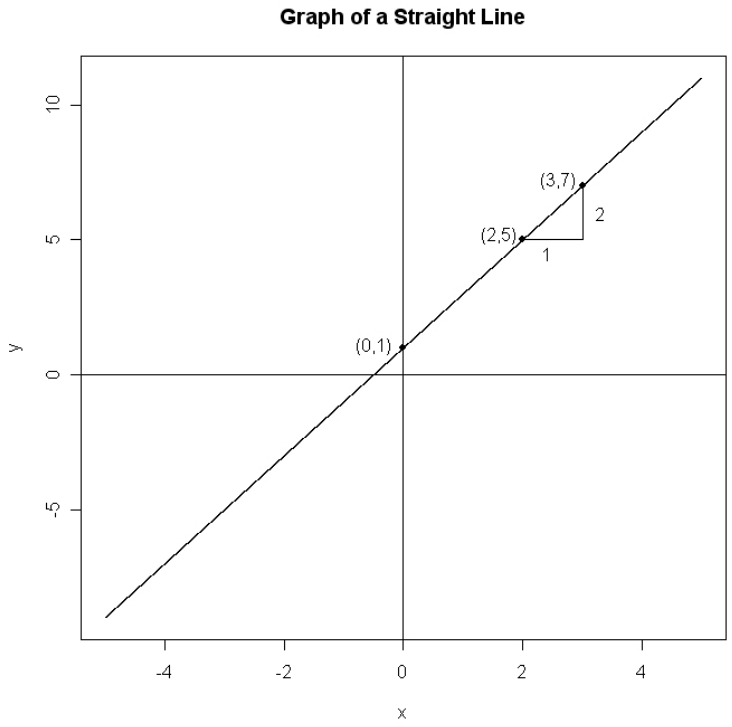


As we will see, this is the function required to predict probabilities of events after running a logistic regression.

(ix) The straight line is the basic function behind linear regression. You may recall that we can define a straight line as

$$f(x) = y = a + bx$$

where a is the intercept of the line (where the line crosses the y-axis), and b is the slope (amount by which y goes up when x increases by one unit). The graph would look like this (letting $a = 1$ and $b = 2$):



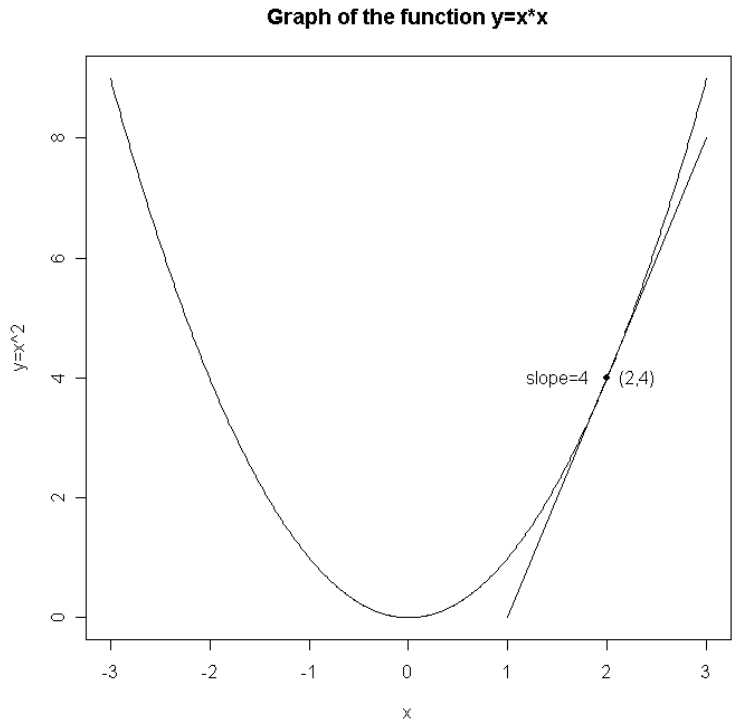
Derivatives: The *derivative* of a function measures the slope of the tangent line to the graph of the function at a given point. For example, if

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

then the derivative is given by

$$f'(x) = 2 \times x.$$

For example, this means that the slope of the tangent line at the point $x = 2$ (with $f(x) = y = 4$) is $2 \times 2 = 4$.



You may recall the following useful facts relating to derivatives:

1. The slope of a line is a measure of how quickly the function is rising or falling as x increases in value.

2. If a function has a maximum or minimum value, then the derivative is usually equal to 0 at that point. In the above, the function has a minimum at $x = 0$, where the value of the derivative is zero.

Derivatives will be used for maximum likelihood estimators, the most common way frequentist estimators are derived.

Integrals: The *indefinite integral* is a synonym for "anti-differentiation". In other words, when we calculate the indefinite integral of a function, we look for a function that when differentiated, returns the function under the integral sign. For example, the indefinite integral of the function $f(x) = x^2$ is given by the

$$\int x^2 \, dx \;=\; \frac{1}{3} \times x^3$$

because the derivative of $\frac{1}{3} \times x^3$ is x^2.

Indefinite integrals are used in many places in statistics. For example, we will see them in the context of regression, when we want to look at the probability density of a regression coefficient. Here, we use an indefinite integral to go from a *joint density* (many variables at once) to a *marginal density* (of a single variable).

The *definite integral* of a function is the area under the graph of that function. This area can be approximated directly from the graph, but exact mathematical formulae are also available from calculus. For example, the area under the the curve ranging from -1 to +2 of the function $f(x) = x^2$ is given by the following definite integral formula:

$$\int_{-1}^{+2} x^2 \, dx \;=\; \frac{1}{3} \times x^3 \Big|_{-1}^{+2} \;=\; \frac{2^3}{3} - \frac{(-1)^3}{3} \;=\; \frac{8}{3} + \frac{1}{3} \;=\; 3.$$

The area under a curve of a probability density function gives the probability of getting values in the region of the definite integral. For example, suppose we wished to calculate the probability that in choosing a random number between 0 and 1 (Uniform density function) the particular number we choose falls between 0.2 and 0.4. This is calculated by the definite integral

$$\int_{0.2}^{0.4} 1 \, dx \;=\; x \Big|_{0.2}^{0.4} \;=\; 0.4 - 0.2 \;=\; 0.2.$$

Aspirin		Tylenol	
Cured	Not Cured	Cured	Not Cured
5	5	5	5
6	4	5	5
6	4	4	6
7	3	4	6
8	2	4	6
8	2	3	7
9	1	3	7
⋮	⋮	⋮	⋮
10	0	0	10

Why is prior information crucial to making any final decisions?

Clearly, *background context* is always important to any conclusions. Can you think of similar medically related examples?


BAYES 'R' US
Specializing In:
Customized complex modelling
Posteriors made to measure
Ask about our optimal design package
We require:
Prior(s)
Data
Model(s)
Design
Our guarantee:
Direct probability statements for your parameters, accounting for current data and your view of past information.
FREQ CORP INC.
P-VALUES
AIC
CONFIDENCE INTERVALS
STEP METHODS
CLEARING SALE ON ALL OUR P-VALUES
SUPPLIED BY CLIENT:
DATA
MODEL(S)
DESIGN
GUARANTEED:
LONG RUN ERROR RATES AVERAGED OVER ALL OUR CUSTOMERS (BUT NO GUARANTEE FOR YOUR DATA).