The Development of a Statistical Computer Software Resource for Medical Research

Thesis submitted in accordance with the requirements of the University of Liverpool

For the degree of Doctor of Medicine

By Iain Edward Buchan

November 2000

Liverpool, England

To my parents…

Preface

Declaration

This thesis is the result of my own work. The material contained in this thesis has not been presented, nor is currently being presented, either wholly or in part for any other degree or other qualification.

Electronic enclosures

This thesis is presented in written and electronic (computer software) forms. The software of this thesis (StatsDirect) can be accessed, from either the enclosed CD-ROM or Internet, as described in Appendix 1. The CD-ROM also contains video clips that relate to the Discussion and Conclusions chapter.

Overview of thesis

Audience

The software part of this thesis is intended for medical researchers, and the written part of this thesis is intended for medical and computational statisticians.

Introduction

The introduction chapter sets out a basic history of statistics and computing, and explores criticisms that medical researchers frequently misuse statistical methods. This situation persists in spite of statistical education in classrooms, textbooks, journal articles and other traditional educational settings. The author presents the need for a new type of resource to support medical researchers in statistical knowledge and computation at the point of research need. The rest of the thesis describes work, the Arcus project, which was carried out by the author in addressing this need.

Methods

The methods chapter outlines the software engineering approach and details the statistical methods used to build the software resource central to this thesis. The fundamental principles of numerical computation are defined and then applied to statistical methods. Mathematical and computational approaches to each statistical method are defined in the light of current literature. The scope of the defined statistical methods maps the statistical needs of medical researchers as indicated by textbooks of medical statistics. In addition to presenting the methods used to establish the resource, the author describes how appropriate development could be sustained over time.

Results

The results chapter is presented in three sections: 1) numerical validation, 2) comparisons with other resources, 3) evidence of use and application to medical research.

Numerical validation

Numerical results of the software are validated against classical examples from the literature. The data used in the examples are contained in the test workbook of the software (StatsDirect) (Appendix 1). The reader can use the software to explore calculations with other data.

Comparisons with other resources

The author compares the software with typical general statistical software resources, and shows that it has greater support of statistical knowledge and orientation to medical research. Software comparisons are summarised in the text of this thesis; the reader should also explore the software and the video clips provided on the enclosed CD-ROM. The integration of statistical knowledge support and calculations can best be demonstrated by using the software, rather than by reading summary observations in the written thesis.

Evidence of use and application to medical research

The author presents peer-reviewed evidence of application of the work of this thesis to medical research. Sustained growth of the uptake of the software is presented as evidence for the likely sustainability of the work into the future.

Discussion and conclusions

The discussion and conclusions chapter examines the work of this thesis against the aim of producing a resource to improve statistical appreciation and practice in medical research. The author discusses key lessons learned through the work. Plans are set out for further research and development on the foundations of this thesis. The author concludes by describing the original contributions of this work to medical research.

Acknowledgements

At the University of Liverpool in the summer of 1989, Dr Peter Bundred encouraged me to develop further a crude prototype of some interactive statistical software that I had written. Dr Bundred and his partner Dr Denise Kitchener have been a great source of inspiration and support to me ever since. I am deeply grateful to them both as respected colleagues and valued friends.

As my MD supervisor and mentor, Professor Richard Edwards has taught me to think critically. A combination of medicine and computing in the early 1990s had few conventional research paths. I am very grateful to Professor Edwards for helping me to pursue research and development without shackles of convention.

In the early years of this work, Professor Peter Armitage answered my niaive letters with grace, and helped me to believe that I could realise a project that many thought too onerous. I am very grateful to Professor Armitage for his kindness and wisdom then and since.

In 1996, inspired by Dr Rudolf Hanka, I moved to a new academic home in Medical Informatics at Cambridge University. Dr Hanka is now both mentor and valued friend, and I have lost count of the times he encouraged me to complete this thesis. Brainstorming with Dr Hanka has taught me to think the unthinkable, with every detail intact.

I am most grateful to Dr Muir Gray and Dr David Pencheon for guiding me to apply my passion for statistics and computing to the health of populations in the field of Public Health. Dr Gray taught me to drive hard at "where the action isn't", and encourages me to think deeply about the management of knowledge.

Dr Norman Pinder, Dr Peter Brambleby and Dr Barry Tennison, have helped me to integrate the writing of this thesis into the academic component of training in

Public Health. I am most grateful for this and to Dr Tennison for his encouragement with and feedback on the software for this thesis.

Over the course of eleven years, many people have inspired me and supported me in the work of this thesis. I am especially grateful to the following people: Mr Jon Honeyball for being my Information Technology guru, patient technical support line, and true friend. Mr Alan Gibbs for his many diligent observations on the software as it evolved, and for his advice on software interfaces for statistical education. Dr Robert Newcombe for educating me on all practical aspects of analysis of binomial proportions and for guiding me with great patience on other topics. Dr Nick Freemantle for his encouragement, rich sets of data and advice on meta-analysis. Dr Fabien Peticolas for his mathematical and computing help with some taxing search algorithms. Mr Christopher West for introducing me to computation for categorical analyses and encouraging me from the earliest phases of this work. Professors Martin Bland and Douglas Altman for answering numerous questions. Dr Pat Altham for her advice in analysis of agreement. Professor Karim Hiriji for his advice on permutational analyses. Professor Douglas Wolfe for his advice on nonparametric confidence intervals. Professor David Hosmer for his advice on goodness of fit tests in logistic regression. Professor Jason Hsu for his advice on methods for multiple comparisons.

I send my deepest gratitude to my dear parents, Evelyn and John Buchan, who have given me their loving support through the long, and sometimes difficult, path leading to this thesis.

Contents

Abstract

The Development of a Statistical Software Resource for Medical Research: MD Thesis of Iain Edward Buchan

Medical research is often weakened by poor statistical practice, and inappropriate use of statistical computer software is part of this problem. The statistical knowledge that medical researchers require has traditionally been gained in both dedicated and ad hoc learning time, often separate from the research processes in which the statistical methods are applied. Computer software, however, can be written to flexibly support statistical practice. The work of this thesis was to explore the possibility of, and if possible, to create, a resource supporting medical researchers in statistical knowledge and calculation at the point of need.

The work was carried out over eleven years, and was directed towards the medical research community in general. Statistical and Software Engineering methods were used to produce a unified statistical computational and knowledge support resource. Mathematically and computationally robust approaches to statistical methods were continually sought from current literature.

The type of evaluation undertaken was formative; this included monitoring uptake of the software and feedback from its users, comparisons with other software, reviews in peer reviewed publications, and testing of results against classical and reference data. Large-scale opportunistic feedback from users of this resource was employed in its continuous improvement.

The software resulting from the work of this thesis is provided herein as supportive evidence. Results of applying the software to classical reference data are shown in the written thesis. The scope and presentation of statistical methods are considered in a comparison of the software with common statistical software resources. This comparison showed that the software written for this thesis more closely matched statistical methods commonly used in medical research, and contained more statistical knowledge support materials. Up to October 31st 2000, uptake of the software was recorded for 5621 separate instances by individuals or institutions. The development has been self-sustaining.

Medical researchers need to have sufficient statistical understanding, just as statistical researchers need to sufficiently understand the nature of data. Statistical software tools may damage statistical practice if they distract attention from statistical goals and tasks, onto the tools themselves. The work of this thesis provides a practical computing framework supporting statistical knowledge and calculation in medical research. This work has shown that sustainable software can be engineered to improve statistical appreciation and practice in ways that are beyond the reach of traditional medical statistical education.

Introduction

"The greatest challenge to any thinker is stating the problem in a way that will allow a solution".

Bertrand Russell (1872 -1970)

Origins of this work

In the summer of 1989, two paradoxical experiences led to the eleven years of research and development underpinning this thesis. First was the author's experience of a clinical research environment at McMaster University Hospital where information technology (IT) was embraced as essential for improving research. At this time, McMaster University Hospital was more advanced in the "Personal Computer (PC) revolution" than many similar centres elsewhere. Second, the author attended a seminar in the UK where a statistician lectured on the danger to research quality of investigators using PCs to perform their own statistical calculations without the expert guidance of a statistician. Both experiences were prophetic, and the paradoxical need to embrace statistical IT without damaging, moreover improving, statistical appreciation, persists today. Here, the author examines how the work of the thesis has addressed this paradox.

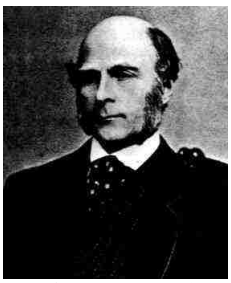
Origins of statistics

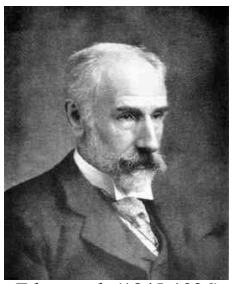
In order to understand the role of statistical computer software in Medical research it is helpful to examine the history of statistical analysis. The statistical methods used in current day medical research are an evolving tapestry of analytical tools, conventions and philosophies. The threads of this tapestry have a much longer history than the discipline that emerged as "statistics" around the turn of the Century (Stigler 1986, MacTutor 2000).

For around one hundred years from the time Legendre first described the least squares method in 1805, different disciplines searched for ways to combine observations in such a way that external variations could be allowed for in

theories about aggregates of data. Astronomers and Geodesists used mathematical theories of gravitation and emerging knowledge of the behaviour of random sums to develop numerical methods for combining observations. The key challenge was to merge these methods with probability theory in a rigorous conceptual framework; a major step forward came in 1809 with the Gauss-Laplace synthesis. Thereafter, Astronomy and Geodesy literature filled with uses of probability in the measurement, comparison and interpretation of uncertainty. Much of this literature used the least squares methods of Legendre to derive coefficients that were used as constants in the development of external theory.

To some extent, psychologists were able to adopt these methods by strict control of experimental conditions, but other researchers, faced with a myriad of potential observations from the largely uncontrollable natural world, needed to develop a different approach to any analysis of combined data dealing with probabilities. The pivotal innovation in this field came with the work of Galton, Edgeworth and Pearson in the late 1800s, later refined by Yule. Their work resulted in generalised regression analysis; this treated errors and conditional probability in such a way that diversity of causes could be reconciled with the ever-present order we observe in the world. Concepts such as correlation had been born along the way and the seeds of many new statistical methods had been sown.


Galton (1822-1911)

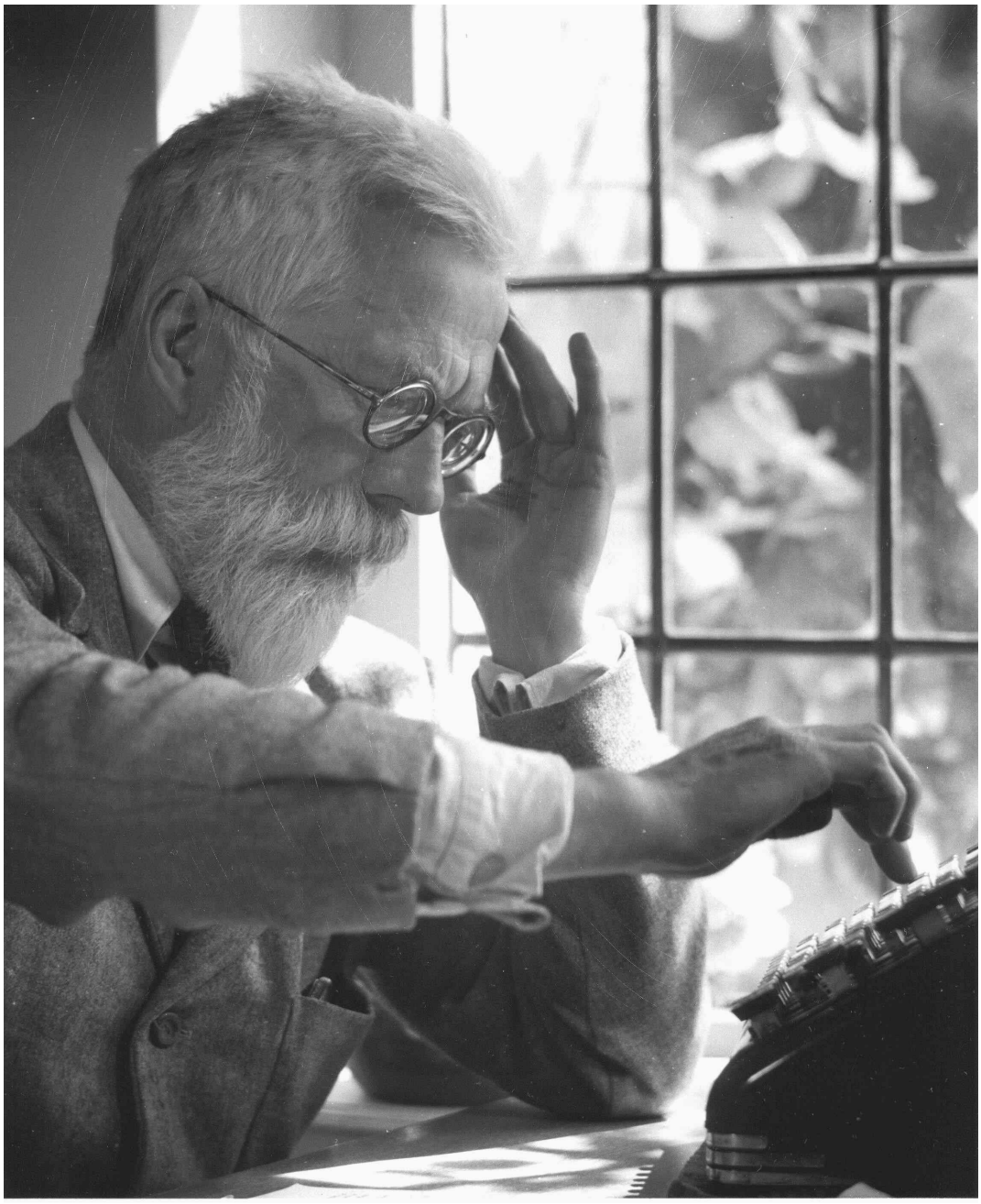

Edgeworth (1845-1926)


Pearson (1857-1936)


Yule (1871-1951)

A couple of decades later Fisher, among his many other accomplishments, perfected regression methods, described analysis of variance and introduced generic concepts such as “maximum likelihood” which underpin a large number of the numerical algorithms in contemporary statistical computer software.

Ronald Fisher is pictured below using a mechanical calculator.



Sir Ronald Aylmer Fisher (1890-1962)

The rise of medical applications of statistics

Medical application of the new science of statistics was largely led by Sir Austin Bradford-Hill, beginning with a series of articles in the journal The Lancet in the late 1930s (Hill, 1937). Hill devoted most of his career to work in this field, and his "Short Text Book of Medical Statistics" was in print for more than fifty years (Hill, 1984; Hill and Hill, 1991). The randomized controlled trial is usually attributed to Hill, and the first widely publicised application of this experimental design was the Medical Research Council streptomycin trial for pulmonary tuberculosis (Medical Research Council, 1948; Vandenbroucke, 1987).

During the 1940s, 1950s and 1960s, the use of formal statistical methods in medical research grew and became a core process in the generation of medical knowledge. The work of Sir Austin Bradford Hill and Sir Richard Doll was an exceptionally influential driver of the adoption of statistical methods in medical research at this time (Vandenbroucke, 1998).

Early medical applications of statistical methods were refined and extended during the last thirty years of the twentieth century. At this time, forums and disciplines, such as clinical epidemiology, evidence-based medicine/health and the Cochrane Collaboration, arose out of the widespread acceptance that statistical methods were essential not only to medical research but also to clinical practice (Evidence-Based Medicine Working Group, 1992; Chalmers et al. 1992).

Current evolution of statistics and its medical applications includes new uses of computer-intensive Bayesian methods, and the incorporation of statistical methods that have grown faster in other fields such as economics (Lilford and Braunholz, 1996; Spiegelhalter et al., 1999).

A brief history of computing machines

The use of calculating machines is an ancient numerical practice. The earliest known calculating machines were tabulating devices such as an abacus. Despite the very crude functionality of such devices, concepts of computing developed far ahead of the practical technology. For example, the twelfth century Tashkent cleric Muhammad ibn Musa Al'Khowarizimi provided the first recorded descriptions of algorithms.

Napier's discovery of logarithms in 1614 spawned the development of analogue scaled devices such as the slide rules that were still in common use in the early 20th century. The first four function mechanical calculator was a stepped drum device invented by Gottfried Leibniz in 1694. The precision required to build such devices was so great that they were not commercially viable until the mid-1800s. Many analogue calculation devices were made, but all required considerable human intervention and left much room for error.

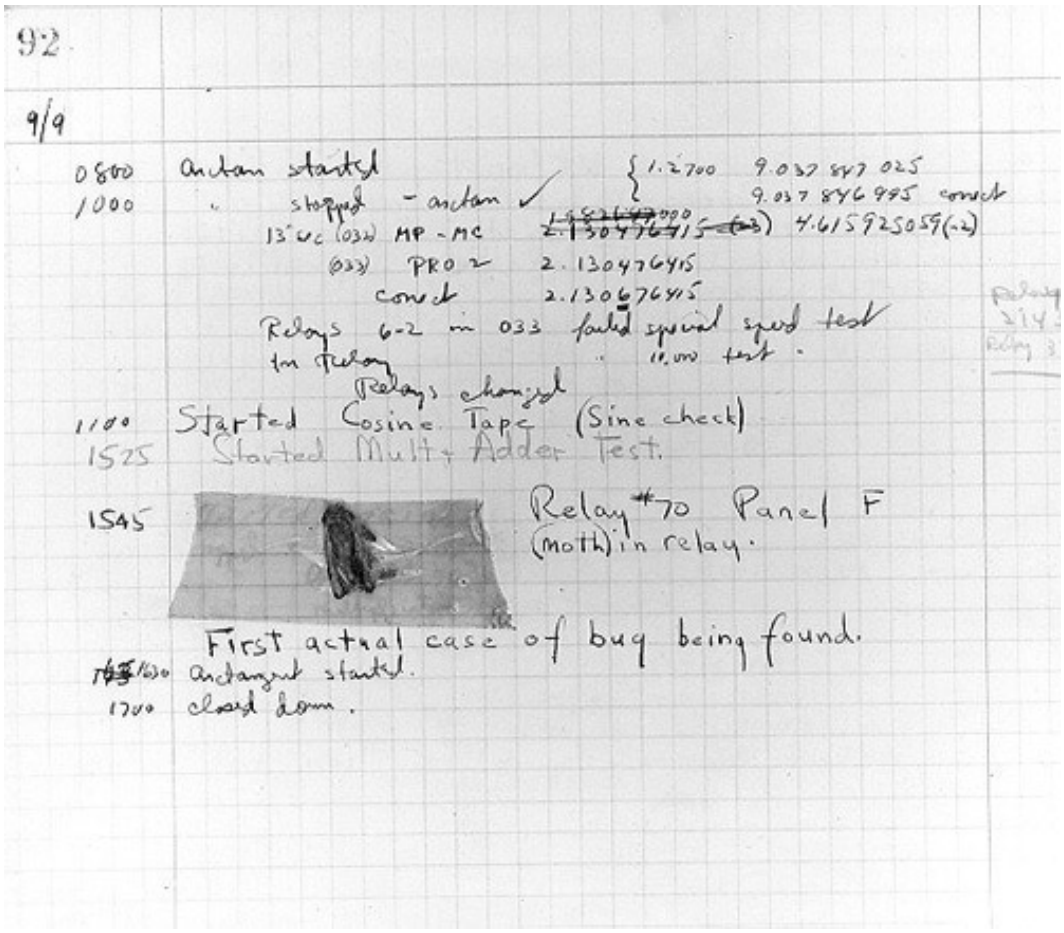
The 1858 Ordinance survey of the British Isles required the solution of 1554 equations with 920 unknowns. It took two teams of "human computers" working independently and in duplicate two and a half years to complete the task (Ernst, 1992). Indeed, Cambridge University Computer Laboratory takes its name from the "computer" staff who worked there long before they were replaced by digital electronic equipment running stored programs (Hanka 1997).

The first digital calculating design was that of Charles Babbage. Babbage built several devices, but his plan of 1837 for a steam powered digital computer was not realised until 1991.

The impetus to create the first practical service computer came with the 1890 US census. Herman Hollerith won the data processing contract and went on to develop "tabulating" equipment as the "Hollerith Tabulating Company" which in 1914 merged to form the "Calculating-Tabulating-Recording Company" later renamed International Business Machines (IBM).

The second great impetus to develop digital computing technology came with the Second World War and the need to decipher enemy encrypted messages. Shortly before this time, Alan Turing made the conceptual leap of "computability" by execution of any describable algorithm on a "universal machine". In 1943, under the guidance of Turing, Colossus (mark I) was completed at Bletchley Park; this ran decryption algorithms that helped defeat Nazi Germany. The following years saw multiple developments toward a stored program computer, first realised in the UK as Manchester (mark I) in 1948.

Programming concepts then developed around real world data processing problems. The team at Cambridge University, including Maurice Wilkes, developed concepts of reusable code such as subroutines. The term "bug" arose in earlier work by Grace Murray Hopper when the moth pictured in the relay pictured here caused an error in a program.



It was not until 1957 that the first compiled "computer language" was realised as FORTRAN (short for formula translator) for the IBM 704 machine. From this time up to the present day, computing equipment and programming have grown vastly in power and complexity. This has been mirrored by a decrease in the level of technological understanding required to operate and to program computers.

As with statistics, growth in *application* of computing has marginalized the mindset of observer-theorist that created it.

Computer-supported numerical reasoning in medical research

For most of the long history of medicine, greater emphasis has been placed upon clinical explanations reasoned through basic science than upon those supported by numerical (statistical and epidemiological) evidence. For example, in his classical descriptions of vibrio, Koch did not acknowledge the earlier epidemiological work of John Snow (Vandenbroucke, 1991).

Influential epidemiological work pre-dates the formation of statistical science; for example, William Farr's work toward the sanitary public health movement (Eyler, 1979). Such work informed the development of medical statistics as a sub-specialist branch of statistics, and many methods currently described as medical statistics are classical epidemiological techniques. The author implies inclusion of epidemiological methods when referring to statistics applied to medical research.

Classical patho-physiological reasoning is deeply ingrained into medical culture, and it gives rise to small-scale, statistically invalid clinical experimentation (Vandenbroucke, 1998; Lewis, 1945). This situation is not necessarily damaging to medical knowledge; on the contrary, innovations of large-scale clinical benefit have been attributed to such experiments (Vandenbroucke, 1998). For these reasons, computer support of numerical reasoning in medical research presents challenges that are different from those in more clearly hypothetico-deductive fields. In other words, medicine has a distinctive epistemology. Software written

to support numerical reasoning in *medical* research should therefore be differently presented to software that is written for general statistical use.

The widespread use of general statistical software in medical research presents two potentially damaging misconceptions. First is the specious acceptance that a large software package is broad enough to support general numerical reasoning in medical research. Second is the use of statistical software as a substitute for consultation with a statistician. When the involvement of a statistician is replaced by inappropriate use of computer software, statistical theory is distanced from observation, and scientific opportunity is lost. Just as medical research is fuelled by clinical observation, advances in statistical science are often born of a marriage of theory and observation. A classical example of an avid observer formulating statistical theory is Francis Galton's conceptual contribution to the development of multiple regression (Stigler, 1986).

Ideally, the knowledge, skills and adaptive reasoning of a statistician would be easily accessible to all medical researchers. The role of statistical software in this situation would be to improve the statistical appreciation of the primary investigator, and thereby support the investigator's experimental reasoning and facilitate his communication with statisticians. Realistically, statisticians are a scarce resource; therefore, a relatively greater role for statistical software is inevitable. Given this reality, the engineers of statistical software should build facilities to trap likely misuse; general statistical software has, however, been criticised for failing to trap misuse in medical research (Altman, 1994). If more engineering effort is put into averting likely misuse of statistical software, then there is an extended opportunity for theorists to seed elements of further enquiry into the minds of observers.

The principal aim of the work of this thesis was to produce a software resource with the potential to reduce the problem of poor statistical appreciation and practice in medical research. The author examines the extent to which this aim has been met by the software presented herein.

Methods

"I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the 'Law of Frequency of Error.' The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along."

Sir Francis Galton (1822-1911)

Software interface development

The guiding principle of interface design was minimisation of software learning in order to maximise contact between the user and statistical knowledge.

In order to minimise software learning the main components (data, report and knowledge objects) were built to function in a very similar way to commonly used spreadsheet, word processor and help system software. In this way the user does not require substantially new software skills when using StatsDirect software for the first time.

The data object was designed to enable the user to manage data for statistical analysis using only skills required to operate common spreadsheet software. The matrix of rows and columns in spreadsheets (more precisely worksheets within workbooks of spreadsheet software) present the user with two main ways of separating groups of data, either to set them out in separate columns or to put them into a single column and create a matching group identifier column. StatsDirect was designed to accommodate both methods of separating groups.

A special interface between the data object and Microsoft Excel spreadsheet software was built because Excel is commonly used to share and manage numerical data in medical research. StatsDirect was designed to put an item into the Excel menu bar that enables the user to push all current data from the Excel workbook they are working on into StatsDirect for statistical analysis.

The report object was designed to enable the user to edit statistical reports using only skills required to operate common word processor software. A verbose style of presentation for results with hooks into context sensitive help was constructed to stimulate statistical appreciation.

A statistical knowledge object was built in the form of an electronic reference book requiring only common hypertext navigation skills. Statistical concepts and methods were presented with worked examples and signposts to further reading and/or advice on seeking help from a statistician.

Grouping of statistical methods by menu items was designed both to relate to the organisation of data analysed and to improve statistical appreciation. For example, the Analysis menu was split into two main parts in order to separate interactive data entry from worksheet-based data entry. The interactive part of the Analysis menu enables the user to perform statistical calculations by entering data as specified on screen by the software. The worksheet-based part of the Analysis menu enables the user to select prepared data from a worksheet for calculation. An example of design for statistical appreciation is the grouping of methods for contingency table analysis; chi-square methods are separated from exact (permutational) methods and the Statistical Method Selection function in the Help menu can assist the user in choosing the appropriate method.

If the user selects a method that is commonly misused then prompts are made as a filter for statistical appreciation. For example, when a two by two chi-square test is selected, the user is asked about the type of study (cohort-study, case control study or neither) that gave rise to the data. Additional calculations are performed

for relative risk if a cohort study is specified or odds ratio if a case-control study is specified, and the user is encouraged to make inference from the confidence intervals around these statistics. When the observed frequencies are small for data entered into a chi-square test, the user is prompted to calculate a Fisher (Fisher-Irwin) exact test.

Software platforms, languages and development tools

The most prevalent computer operating systems were chosen as targets for software development. The first wave of Arcus software (1990-5) ran under IBM and Microsoft Disk Operating System (DOS), the second wave (1995-9) ran on Microsoft's 16 bit Windows environment and the current StatsDirect software runs on Microsoft's 32 bit Windows environments. At each period in the software developments of this thesis, the target operating system was chosen as that which a medical researcher would be likely to have access to at work or home.

BASIC and FORTRAN were used as the main development languages. These languages have different strengths and a similar syntax. C and Assembly Language were used for a small number of low-level system routines.

Historically, the statistical community has used FORTRAN to communicate algorithms in scientific literature and to build statistical software. Although not useful for systems level programming, FORTRAN remains strongly supported for writing and optimally compiling numerical algorithms. Compaq Visual FORTRAN was used as the development tool for the most computationally intensive numerical algorithms in StatsDirect. Prior to this, Microsoft FORTRAN was used for parts of the 16 bit Windows and DOS editions of Arcus software.

Enhanced versions of the BASIC language are used in popular Rapid Application Development (RAD) environments such as Microsoft Visual Basic. The Visual Basic 6 RAD of the Microsoft Visual Studio development environment was used for StatsDirect development for a number of reasons, including ease of interface design and interactive debugging of algorithms. Prior to this, Microsoft Basic

Professional Development System and 16 bit versions of Microsoft Visual Basic were used for the front end of 16 bit Windows and DOS editions of Arcus software.

Two pre-built software components are used in StatsDirect, these are the Tidestone Formula One spreadsheet object and the Tidestone First Impression charting object (Tidestone Corporation, 1999). In order to maximise speed of sorting arrays, the Stamina library of assembly language routines was employed for sorting (Microdexterity Corporation, 1999). All other components were written by the author using the Microsoft Visual Studio development environments (Microsoft Corporation, 1998).

Numerical precision and error

"Although this may seem a paradox, all exact science is dominated by the idea of approximation."
Russell, Bertrand (1872-1970)

Numbers with fractional parts (real/floating-point as opposed to integer/fixed-point numbers) cannot all be fully represented in binary computers because computers cannot hold an infinite number of bits (binary digits) after the decimal point. The only real numbers that are represented exactly are those that can be expressed as a fraction with denominator that is a power of two (e.g. 0.25); just as the only terminating (finite) decimals are those expressible as a fraction with denominator that is a power of ten (e.g. 0.1). Many real numbers, one third for example, cannot be expressed as a terminating decimal or binary number. Binary computers therefore represent many real numbers in approximate form only, the global standard for doing this is IEEE Standard Floating-Point Representation (IEEE, 1985).

Numerical algorithms written in Microsoft Visual Basic and Compaq Visual FORTRAN comply with both single (32 bit) and double (64 bit) precision IEEE Standard Floating-Point Representation (Microsoft Corporation, 1998; Compaq

Corporation, 2000; IEEE, 1985). All real numbers in StatsDirect are handled in double precision.

Arithmetic among floating point numbers is subject to error. The smallest floating point number which, when added to 1.0, produces a floating-point number different to 1.0 is termed the machine accuracy ε_m (Press et al., 1992). In IEEE double precision ε_m is approximately 2.22×10^{-16}. Most arithmetic operations among floating point numbers produce a so-called round-off error of at least ε_m. Some round-off errors are characteristically large, for example the subtraction of two almost equal numbers. Round-off errors in a series of arithmetic operations seldom occur randomly up and down. Large round-off error at the beginning of a series of calculations can become magnified such that the result of the series is substantially imprecise, a condition known as instability. Algorithms in StatsDirect were assessed for likely causes of instability and common stabilising techniques, such as leaving division as late as possible in calculations, were employed.

Another error inherent to numerical algorithms is the error associated with approximation of functions; this is termed truncation error (Press et al., 1992). For example, integration is usually performed by calculating a function at a large discrete number of points, the difference between the solution obtained in this practical manner and the true solution obtained by considering every possible point is the truncation error. Most of the literature on numerical algorithms is concerned with minimisation of truncation error. For each function approximation in StatsDirect, the most precise algorithms practicable were written in the light of relevant, current literature.

Evaluating arithmetic expressions

A set of routines was written to evaluate arithmetic expressions entered in text form. The text expression is first scanned to replace constants; it is then split into sections separated by brackets and/or operators. Functions are evaluated for each bracketed section and nests thereof. Each step of the solution is calculated in double (64 bit) precision.

In StatsDirect, the arithmetic engine is used to form an algebraic calculator and a user-defined transformation engine for worksheet based data. The following constants, functions and operators are supported:

Constants

PI	3.14159265358979323846 (π)
EE	2.71828182845904523536 (e)

Arithmetic functions

ABS	absolute value
CLOG	common (base 10) logarithm
CEXP	anti-log (base 10)
EXP	anti-log (base e)
LOG	natural (base e) logarithm
LOGIT	logit: log(p/[1-p]), p=proportion
ALOGIT	anti-logit: exp(l)/[1+exp(l)], l=logit
SQR	square root
!	factorial (maximum 170)
LOG!	log factorial
IZ	normal deviate for a p value
UZ	upper tail p for a normal deviate
LZ	lower tail p for a normal deviate

The largest factorial allowed is 170! but larger factorials can be worked with by using logarithms, Log factorials are supported via the LOG! function, e.g. LOG!(171).

Arithmetic operators

^	exponentiation (to the power of)
+	addition
-	subtraction
*	multiplication
/	division
\	integer division

Operator precedence

Calculations give an order of priority to arithmetic operators. For example, the result of the expression "6 - 3/2" is 4.5 and not 1.5 because division takes priority over subtraction.

Priority of arithmetic operators in descending order:
Exponentiation (^)
Negation (-X) (Exception = x^-y; i.e. 4^-2 is 0.0625 and not -16)
Multiplication and Division (*, /)
Integer Division (\)
Addition and Subtraction (+, -)

Trigonometric functions

ARCCOS	arc cosine
ARCCOSH	arc hyperbolic cosine
ARCCOT	arc cotangent
ARCCOTH	arc hyperbolic cotangent
ARCCSC	arc cosecant

ARCCSCH	arc hyperbolic cosecant
ARCTANH	arc hyperbolic tangent
ARCSEC	arc secant
ARCSECH	arc hyperbolic secant
ARCSIN	arc sine
ARCSINH	arc hyperbolic sine
ATN	arc tangent
COS	cosine
COT	cotangent
COTH	hyperbolic cotangent
CSC	cosecant
CSCH	hyperbolic cosecant
SIN	sine
SINH	hyperbolic sine
SECH	hyperbolic secant
SEC	secant
TAN	tangent
TANH	hyperbolic tangent

To convert degrees to radians, multiply degrees by $\pi/180$.

To convert radians to degrees, multiply radians by $180/\pi$.

Logical functions

AND	logical AND
NOT	logical NOT
OR	logical OR
<	less than
=	equal to
>	greater than

Counting and grouping

Categorical data are usually recorded in individual record form and analysed in aggregate form (Agresti, 1996). Database software may be used to record and encode the raw data. Spreadsheet software may be used to investigate the data exported from databases in row (record) and column (field) format.

Statistical analysis of categorical data usually requires comparison of counts (frequencies) of observations in different categories (Agresti, 1996). Routines were written to count raw and cumulative frequencies of different observations in a single variable. For the cross classification of observations from two variables, a cross tabulation routine was written. As two variable cross tabulation produces a two way contingency table, statistical methods appropriate to the dimension of the resulting two way table are linked to this routine (with appropriate prompts to the user).

Different groups of data can be arranged into separate columns of a worksheet or they can be arranged into a single "data" column with a matching "group identifier" column. The latter form of grouping is common in statistical software because it is close to the database format in which raw data are frequently collected. The former method (separate columns), however, is the common form of presentation of different groups in textbooks (Armitage and Berry, 1996; Altman 1991; Bland, 1996). In order to address both statistical appreciation and ease of data management, StatsDirect was written to handle both types of data grouping and to translate between them (see example below):

Group ID	Data	<--->	Data~G1	Data~G2	Data~G3
1	1.1		1.1	0.7	1.9
1	1.3		1.3	1.0	2.2
1	0.9		0.9	0.6	1.7
2	0.7		1.5	1.1	
2	1.0		1.3		
3	1.9				
1	1.5				
1	1.3				
2	0.6				
3	2.2				
3	1.7				
2	1.1				

Another re-grouping of data is rotation of a block of cells in a worksheet such that rows and columns are transposed. A block rotation facility was included in the Data menu facilities of StatsDirect.

Searching and translation of dates and text

Worksheet data may contain numbers, dates or text. Spreadsheet software offers a wide range of facilities for managing such data in rows and columns. Beyond conventional spreadsheet facilities, a set of routines to search and optionally replace or remove worksheet data was written for StatsDirect.

The search facility enables the user to limit the search to a condition (e.g. greater than or equal to 1). Data that match the search condition are counted, replaced with a specified value or removed. If data to be removed are paired with neighbouring columns then the user has the option to remove the entire row where a match to the search condition is found. The removal or replacement of data may be important where data have been exported from a recording system that uses a specific value to mark missing observations. The statistical routines in StatsDirect treat empty cells as missing data and use the value $3.0 \times 10^{+300}$ to represent them internally where methods take account of missing observations in calculation.

All dates are handled in a conventional spreadsheet manner. In order to make sure that the user fully expands the data they use in calculations, methods that handle dates (specifically survival analysis) request the input of time intervals and not dates. A data manipulation routine is provided for the calculation of time/date intervals from a reference time/date. Standard worksheet functions (included in StatsDirect) can be used to manipulate date data adequately for the purpose of data preparation.

Problems of translation between text and numbers can occur when observations are coded differently for data management and analysis. StatsDirect was designed to encode text as numbers, optionally via a translation table.

Sorting, ranking and normal scores

Many statistical calculations require data to be sorted. Fast execution of sorting was achieved using a hybrid Quicksort-Shell sort algorithm written at very low level (Machine Assembly Language). Quicksort is the fastest known sort algorithm for most situations; Shell's method is added to cope with notable exceptions when numbers are small or many of the data are pre-sorted (Knuth 1998; Microdexterity Corporation, 1999).

Ranking is effectively a transformation that pulls in both tails of a distribution. Statistical methods based upon ranks are therefore useful for inference that does not depend upon the data being from a particular distribution, i.e. nonparametric data. Most nonparametric methods described elsewhere in this thesis use a ranking algorithm based upon the list-merge method (Knuth, 1998). For ranking equivalent (tied) observations, the average value of the order statistics for tied values is taken as the rank. StatsDirect gives the option to calculate various summary statistics for ties; these were chosen as a selection of tie-adjustments used in nonparametric methods (Conover, 1999; Hollander and Wolfe, 1999).

Van der Waerden's and Blom's methods are used to normalise ranks in order to calculate approximate normal scores (Conover, 1999; Altman, 1991).

Van der Waerden:

$$s = \Phi\left(\frac{r}{n+1}\right)$$

- where s is the normal score for an observation, r is the rank for that observation, n is the sample size and $\Phi(p)$ is the p^{th} quantile from the standard normal distribution.

Blom:

$$s = \Phi\left(\frac{r - 3/8}{n + 1/4}\right)$$

- where s is the normal score for an observation, r is the rank for that observation, n is the sample size and $\Phi(p)$ is the p^{th} quantile from the standard normal distribution.

Expected normal order scores are calculated as (David, 1981; Royston, 1982; Harter, 1961):

$$s = \frac{n!}{(r-1)!(n-r)!}\int_{-\infty}^{\infty} x\{1 - \Phi(x)\}^{r-1}\{\Phi(x)\}^{n-r}\phi(x)dx$$

$$\phi(x) = \frac{1}{\sqrt{2\pi}\exp(-0.5x^2)}$$

- where s is the normal score for an observation, r is the rank for that observation, n is the sample size, $\phi(p)$ is the standard normal density for p and $\Phi(p)$ is the p^{th} quantile from the standard normal distribution. The solution is found by numerical integration.

Pairwise calculations

The intermediate steps of many nonparametric calculations require pairwise enumeration of data (Conover, 1999; Hollander and Wolfe 1999). The common pairwise scenarios covered here are differences (contrasts), means and slopes.

Pairwise differences

Given two variables X and Y, every possible contrast between each X value and each Y value is made. If X and Y consist of n and m observations respectively then there are nm possible contrasts (D):

$$D_{k\ (k = 1 \text{ to } nm)} = [X_i - Y_j]_{\ (i=1 \text{ to } n,\ j = 1 \text{ to } m)}$$

An example of application of this method is the construction of a confidence interval for the difference between two independent means or medians (Conover, 1999).

Pairwise means

Given a variables X, every possible pairwise mean within X is calculated. If X consists of n observations then there are n(n+1)/2 possible pairwise means (M):

$$M_{k\ (k = 1 \text{ to } n(n+1)/2)} = [(X_i + X_j)/2]_{\ (i = 1 \text{ to } n,\ j=i \text{ to } n)}$$

An example of application of this method is the construction of a confidence interval for the median difference for a pair of samples (Conover, 1999).

Pairwise slopes

Given a response variable Y and a predictor variable X, every feasible pairwise slope between XY pairs is calculated. If each of X and Y consist of n observations then there are n(n-1)/2 possible pairwise slopes (S):

$$S_{k\ (k = 1 \text{ to } n(n-1)/2)} = [(Y_i - Y_j)/(X_i - X_j)]_{\ (i = 1 \text{ to } n,\ j = i \text{ to } n,\ i <> j)}$$

An example of application of this method is Theil type nonparametric linear regression that estimates the slope as the median of all pairwise slopes. A confidence interval based upon Kendall's τ distribution is given for the median slope (Conover, 1999).

Transformations

A range of pre-defined transformation functions and a general transformation based upon the arithmetic expression evaluator were written.

The subject of transformation can be especially confusing for the non-statistician for two main reasons. First, transformations are performed for different purposes. Second, transformation changes the measurement scale of data; therefore, there is considerable scope to misconceive the use of transformed data (Bland and Altman, 1996c; Armitage and Berry, 1994). Additional explanation on transforming data was thus added to the help system.

Logarithmic

Natural (base e) and common (base 10) log transform functions were written. The user is given the option to allow substitution of indeterminable natural logs (values less than or equal to 0) with the inverse hyperbolic sine value. Log transform stabilizes variance, linearizes increasing slopes in X in relation to another variable and normalizes positively skewed distributions.

Logit

Logit is defined as the natural log of p/(1-p) where p is a proportion. Indeterminable values (p equal to 0 or 1) are marked as missing. Logit transform linearizes certain sigmoid distributions of proportions.

Probit

Probit is defined as 5 + the 1-p quantile from the standard normal distribution, where p is a proportion. Indeterminable values (p equal to 0 or 1) are marked as missing. Probit transform linearizes certain sigmoid distributions of proportions.

Angular

Angle is defined as inverse sine of the square root of a proportion. Angular transform linearizes certain sigmoid distributions of proportions and stabilizes their variance.

Cumulative

Cumulative transform is the sequential addition of a series of data that have some meaningful order. Data are transformed from individual observations into the cumulative set.

Ladder of powers

The term ladder of powers is used to describe a series of transformations that have increasing power to pull in the right hand tail of a distribution. The ladder consists of:

$1/x^2$	reciprocal of square (power -2)
$1/x$	reciprocal (power -1)
log(x)	natural logarithm
$x^{1/2}$	square root (power 0.5)
x^2	square (power 2)

P values and confidence intervals

Calculated probability or P values are given too much emphasis and misused in Medical research (Altman, 1994; Gardener and Altman, 1989).

Inappropriate presentation of P values includes the use of a large number of decimal places and quotation of zero probabilities (many statistical software packages display P of 0.000 implying or stating P = 0.000 when P < 0.000). StatsDirect displays a default four decimal places (user can re-set this to another value) for P values and uses the colour green for displaying P values in reports. P of less than the minimum number displayed as P < 0.0…1 and P of greater than the maximum number is displayed as P > 0.9…9. The help system uses the convention of inference as "statistically highly significant" for P < 0.001.

Confidence intervals were given greater prominence than P values by putting them at the end of results from methods, i.e. the bottom line on a calculation is the confidence interval. With methods that are usually associated with P values and not with confidence intervals (e.g. two by two chi-square test), the most appropriate estimate of effect is presented with a confidence interval at the end of the result. Encouragement to use confidence intervals in this way sometimes necessitates additional interaction with the user after a method is selected; here a trade-off between convenience and statistical appreciation is made in favour of the latter. The default level for confidence intervals is set at 95%; this can be re-set by the user.

Many papers, books and computer software packages use unnecessarily conservative or unstable approximation formulae designed for use when computing power was not readily accessible (Newcombe, 1998a, 1998b, 1998c; McCullough and Wilson, 1999). StatsDirect uses the most robust and reliable algorithms for calculation of P values and confidence intervals. Exact methods were used wherever practicable (in terms of computer memory consumed and time taken to calculate the result) in order to maximise accuracy, and thus robustness. The term exact implies a theoretical truncation error close to zero.

Probability distributions

Algorithms to calculate areas and quantiles for common probability distributions were written with the aim of minimising truncation error within practical computing limits. Academic literature was searched for computational methods. Explanation of probability distributions in the help system was written to assist the user in practical concepts of applied probability.

Normal

Distribution function, Φ(z), of a standard normal variable z:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t/2} dt$$

StatsDirect calculates Φ(z) from the complement of the error function (errc):

$$\Phi(z) = errc\left(\frac{-z}{\sqrt{2}}\right) \Big/ 2$$

$$errc(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2} dt$$

The tail area of the normal distribution is evaluated to 15 decimal places of accuracy using the complement of the error function (Abramowitz and Stegun, 1972; Johnson and Kotz, 1970). The quantiles of the normal distribution are calculated to 15 decimal places using a method based upon AS 241 (Wichura 1988).

Chi-square

The distribution function F(x) of a chi-square random variable x with *n* degrees of freedom is:

$$F(x) = \frac{1}{2^{n/2}\Gamma(n/2)}\int_0^x t^{n/2-1}e^{-t/2}dt$$

Γ(*) is the gamma function:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt\ ,\ \text{x>0}$$

The probability associated with a chi-square random variable with *n* degrees of freedom is calculated. A reliable approach to the incomplete gamma integral is used (Shea, 1988). Chi-square quantiles are calculated for *n* degrees of freedom and a given probability using the Taylor series expansion of Best and Roberts (1975) when P ≤ 0.999998 and P ≥ 0.000002, otherwise a root finding algorithm is applied to the incomplete gamma integral.

Student's t

The distribution function of a t distribution with n degrees of freedom is:

$$f(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\Gamma(n/2)}\int_{-\infty}^t \left(1+\frac{x^2}{n}\right)^{-(n+1)/2} dx$$

Γ(*) is the gamma function:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt\ ,\ \text{x > 0}$$

A t variable with n degrees of freedom can be transformed to an F variable with 1 and n degrees of freedom as $t^2=F$. An F variable with ν_1 and ν_2 degrees of freedom can be transformed to a beta variable with parameters $p=\nu_1/2$ and $q=\nu_2/2$ as beta= $\nu_1 F(\nu_2+\nu_1 F)$. The beta distribution with parameters p and q is:

$$beta(x) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}\int_0^x t^{p-1}(1-t)^{q-1}dt$$

The relationship between Student's t and the beta distribution is used to calculate tail areas and percentage points for t distributions. Soper's reduction method is used to integrate the incomplete beta function (Majumder and Bhttacharjee, 1973a, 1973b; Morris, 1992). A hybrid of conventional root finding methods is used to invert the function. Conventional root finding proved more reliable than some other methods such as Newton Raphson iteration on Wilson Hilferty starting estimates (Berry et al., 1990; Cran et al., 1977). StatsDirect does not use the beta distribution for the calculation of t percentage points when there are more than 60 degrees of freedom (n), here a less computationally demanding approximation is reliable (Hill 1970). When n is 2, t is calculated as sqr(2/(P(2-P))-2). When n is 1, t is calculated as cos((Pπ)/2)/sin((Pπ)/2).

F (variance ratio)

An F variable with ν_1 and ν_2 degrees of freedom can be transformed to a beta variable with parameters $p=\nu_1/2$ and $q=\nu_2/2$ as $beta=\nu_1 F(\nu_2+\nu_1 F)$. The beta distribution with parameters p and q is:

$$beta(x)=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}\int_0^x t^{p-1}(1-t)^{q-1}dt$$

Γ(*) is the gamma function:

$$\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt\ ,\ x>0$$

Tail areas and percentage points are calculated for given numerator and denominator degrees of freedom. Soper's reduction method is used to integrate the incomplete beta function (Majumder and Bhttacharjee, 1973a, 1973b; Morris, 1992). A hybrid of conventional root finding methods is used to invert the function. Conventional root finding proved more reliable than some other methods such as Newton Raphson iteration on Wilson Hilferty starting estimates (Berry et al., 1990; Cran et al., 1977).

Studentized range (Q)

The Studentized range, Q, is a statistic due to Newman (1939) and Keuls (1952) that is used in multiple comparison methods. Q is defined as the range of means divided by the estimated standard error of the mean for a set of samples being compared. The estimated standard error of the mean for a group of samples is usually derived from analysis of variance.

Tail areas and percentage points are calculated for a given number of samples and sample sizes using the method of Copenhaver and Holland (1988). Other commonly cited methods produce a smaller range of results and are less precise (Gleason, 1999; Lund and Lund, 1983; Royston, 1987).

Spearman's rho

For two rankings ($x_1,x_2...x_n$ and $y_1,y_2...y_n$) of n objects without ties:

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

T is related to the Spearman rank correlation coefficient (ρ) by:

$$\rho = 1 - \frac{6T}{n^3 - n}$$

The probability of obtaining a value ≥ T is calculated for the Hotelling-Pabst test statistic (T) or Spearman's rho (ρ) by summation across all permutations when n < 10 and by an Edgeworth series approximation when n ≥ 10 (Best and Roberts, 1975; Hotelling and Pabst, 1936). The Edgeworth series results for n ≥ 10 are accurate to at least four decimal places.

The inverse is calculated by finding the largest value of T that gives the calculated upper tail probability (using the methods above) closest to but not less than the P value entered.

Kendall's tau

Consider two samples, x and y, each of size n. The total number of possible pairings of x with y observations is n(n-1)/2. Now consider ordering the pairs by the x values and then by the y values. If $x_3 > y_3$ when ordered on both x and y then the third pair is concordant, otherwise the third pair is discordant. S is the difference between the number of concordant (ordered in the same way, n_c) and discordant (ordered differently, n_d) pairs.

Tau (τ) is related to S by:

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

If there are tied (same value) observations then τ_b is used:

$$\tau_b = \frac{S}{\sqrt{\left[n(n-1)/2 - \sum_{i=1}^{t} t_i(t_i-1)/2\right]\left[n(n-1)/2 - \sum_{i=1}^{u} u_i(u_i-1)/2\right]}}$$

- where t_i is the number of observations tied at a particular rank of x and u is the number tied at a rank of y. When there are no ties $\tau_b = \tau$.

Calculation of probability for τ_b requires too many computations for routine use.

Probabilities for S are calculated by summation across all permutations when $n \leq 50$ or by and an Edgeworth series approximation when $n > 50$ (Best and Gibbs, 1974). Samples are assumed to have been ranked without ties.

The inverse is calculated by finding the largest value of S that gives the calculated upper tail probability (using the methods above) closest to, but not less than, the P value entered. The inverse of Kendall's statistic is calculated more accurately than Best's widely quoted 1974 table.

Binomial

A binomial distribution occurs when there are only two mutually exclusive possible outcomes, for example the outcome of tossing a coin is heads or tails. It is usual to refer to one outcome as "success" and the other outcome as "failure". The binomial distribution can be used to determine the probability, P(r) of exactly r successes:

$$P(r) = \binom{n}{r} p^r (1-p)^{n-r}$$

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{(n-r)}$$

Here p is the probability of success on each trial.

In many situations, the probability of interest is not that associated with *exactly* r successes but instead it is the probability of *r or more* (≥ r) or *at most r* (≤ r) successes. Here the cumulative probability is calculated:

$$P(\geq r) = \sum_{i=r}^{n} P(r_i)$$

$$P(\leq r) = \sum_{i=1}^{r} P(r_i)$$

Probability for exactly r and the cumulative probability for (≥, ≤) r successes in n trials are calculated. The gamma function is a generalised factorial function and it is used to calculate each binomial probability. The core algorithm evaluates the logarithm of the gamma function (Cody and Hillstrom, 1967; Abramowitz and Stegun 1972; Macleod, 1989) to the limit of 64-bit precision.

Γ(*) is the gamma function:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt\,, \quad x>0$$

Γ(1)=1

Γ(x+1)=xΓ(x)

Γ(n)=(n-1)!

Poisson

Both the mean and variance of a Poisson distribution are equal to μ. The probability of r events happening in unit time with an event rate of μ is:

$$P(r) = \frac{e^{-\mu}\mu^{r}}{r!}$$

The summation of this Poisson frequency function from zero to r will always be equal to one as:

$$e^{\mu} = \sum_{r=0}^{\infty}(\mu^{r} / r!)$$

StatsDirect calculates cumulative probabilities that (≤, ≥, =) r random events are contained in an interval when the average number of such events per interval is μ. The gamma function is a generalised factorial function and it is used to calculate each Poisson probability (Knusel, 1986). The core algorithm evaluates the logarithm of the gamma function (Cody and Hillstrom, 1967; Abramowitz and Stegun 1972; Macleod, 1989) to the limit of 64-bit precision.

Γ (*) is the gamma function:

$$\Gamma(x) = \int_{0}^{\infty} t^{x-1}e^{-t}dt\ ,\ \text{x>0}$$

Γ(1)=1

Γ(x+1)=xΓ(x)

Γ (n)=(n-1)!

Non-central t

Non-central t (T) represents a family of distributions that are shaped by ν degrees of freedom and a non-centrality parameter (δ).

Non-central t may be expressed in terms of a normal and a chi-square distribution:

$$T = \frac{z}{\sqrt{\chi^2 / \nu}}$$

- where z is a normal variable with mean δ and variance 1 and χ^2 is a chi-square random variable with degrees of freedom (Owen, 1965).

StatsDirect evaluates the cumulative probability that a t random variable is less than or equal to a given value of T with ν degrees of freedom and non-centrality parameter δ (Lenth, 1989; Owen, 1965; Young and Minder, 1974; Thomas, 1979; Chou, 1985; Boys 1989; Geodhart and Jansen, 1992). The inverse of T is found by conventional root finding methods to six decimal places of accuracy.

Sample sizes

Sample sizes necessary to avoid given levels of type II error are estimated. The methods covered are comparison of means using Student t tests, comparison of proportions and population surveys.

Given the scope for misinterpretation and misuse of power and sample size calculations, supporting explanation was put into the help system. Help content was grouped with other guidance on study design and analysis in order to encourage users to think more about design issues *before* starting studies, the converse is frequently observed (Altman, 1994).

The following definitions are used in each of the sample size estimates:

- Power is the probability of detecting a real effect.
- Alpha is the probability (two sided) of detecting a false effect and is equal to 1- confidence level.
- Calculated sample size n is rounded up to the nearest integer.

Population survey

The minimum number of subjects required for a survey of a population is estimated for a specified difference in the proportion of individuals in that population that display the factor under study (Colton, 1974).

The estimated sample size n is calculated as:

$$n = \frac{sn}{1 + sn / N}$$

$$sn = \frac{z^2 p(1-p)}{d^2}$$

- where p is an estimate of the rate at which factor occurs in the population, d is deviation from p that you would tolerate (i.e. p ± d), and z is a quantile from the standard normal distribution for a two tailed probability alpha.

Paired cohort

The minimum number of subject pairs required to detect a specified relative risk is estimated for a given power and alpha (Dupont, 1990; Breslow and Day, 1980).

The estimated sample size n is calculated as:

$$n = \frac{\left[\frac{Z_{\alpha/2}}{2} + Z_{\beta}\sqrt{p_a(1-p_a)}\right]^2}{(p_a - 0.5)^2 (p_x p_y)}$$

$$p_a = \frac{p_y}{p_x + p_y}$$

$$p_y = p_1(1-p_0) - r * \sqrt{p_1(1-p_1)p_0(1-p_0)}$$

$$p_x = p_0(1-p_1) - r * \sqrt{p_1(1-p_1)p_0(1-p_0)}$$

- where α is alpha, β is 1 - power, Z_p is the standard normal deviate for probability p, r is the correlation coefficient for failure between paired subjects, P_0 is the event rate in the control group, P_1 is the event rate in the experimental group and the relative risk is P_1/P_0.

Independent cohort

The minimum number of case subjects required to detect a specified relative risk or experimental event rate is estimated for a given power and alpha. The sample size is also given as a continuity corrected value intended for use with corrected chi-square and Fisher's exact tests (Casagrande et al., 1978; Meinert 1986; Fleiss, 1981; Dupont, 1990).

The estimated sample size n is calculated as:

$$n = \frac{\left[Z_{\alpha}\sqrt{(1+1/m)\bar{p}(1-\bar{p})} + Z_{\beta}\sqrt{p_0(1-p_0)/m + p_1(1-p_1)}\right]^2}{(p_0 - p_1)^2}$$

$$\bar{p} = \frac{p_1 + m\,p_0}{m+1}$$

$$n_c = \frac{n}{4}\left(1+\sqrt{1\frac{2(m+1)}{n\,m|p_0-p_1|}}\right)^2$$

- where α is alpha, β is 1 - power, n_c is the continuity corrected sample size and Z_p is the standard normal deviate for a probability p, m is the number of control subjects per case, p_0 is the event rate in the control group, p_1 is the event rate in the experimental group and the relative risk is p_1/p_0.

Matched case-control

The minimum sample size necessary to detect a specified odds ratio OR is estimated for a given power and alpha. If there is more than one control per case then the reduction in sample size relative to a paired study that can be obtained using m controls per case is also calculated (Dupont, 1988).

The estimated sample size n is calculated as:

$$n = \frac{\left[(1/\sigma_\psi)Z_{\alpha/2} + Z_\beta\right]^2}{d^2}$$

$$\sigma_\psi = \sqrt{\sum_{k=1}^{m}\frac{k\,t_k\psi(m-k+1)}{(k\psi+m-k+1)^2}}$$

$$t_k = p_1(k-1)p_{0+}^{k-1}(1-p_{0+})^{m-k+1} + (1-p_1)\,k\,p_{o-}^{k} - (1-p_{0-})^{m-k}$$

$$p_{0+} = \frac{p_1p_0 + r\sqrt{p_1(1-p_1)p_0(1-p_0)}}{p_1}$$

$$p_{0-} = \frac{p_0(1-p_1) - r\sqrt{p_1(1-p_1)p_0(1-p_0)}}{1-p_1}$$

$$d = \frac{\left[\sum_{k=1}^{m}\frac{k\,t_k\,\psi}{k\psi+m-k+1}\right]-1}{\sigma_\psi}$$

- where α is alpha, β is 1 - power, , ψ is the odds ratio, Z_p is the standard normal deviate for probability p, r is the correlation coefficient for failure between paired subjects, p_0 is the probability of exposure in the control group, p_1 is the probability of exposure in the experimental group and the relative risk is p_1/p_0.

Independent case-control

The minimum number of case subjects required to detect a specified odds ratio or case exposure rate is estimated for a given power and alpha. The sample size is also given with a continuity correction for use with corrected chi-square and Fisher's exact tests (Schlesselman, 1982; Casagrande et al. 1978; Dupont, 1990).

The estimated sample size n is calculated as:

$$n = \frac{\left[Z_\alpha \sqrt{(1+m)\overline{p}'(1-\overline{p}')} + Z_\beta \sqrt{p_1(1-p_1) + m\,p_0(1-p_0)}\right]^2}{(p_1 - p_0)^2}$$

$$\overline{p}' = \frac{p_1 + p_0 / m}{1 + 1/m}$$

$$p_1 = \frac{p_0 \psi}{1 + p_0(\psi - 1)}$$

$$n_c = \frac{n}{4}\left(1 + \sqrt{1 \frac{2(m+1)}{n\,m\,|p_0 - p_1|}}\right)^2$$

- where α is alpha, β is 1 - power, ψ is the odds ratio, Z_p is the standard normal deviate for probability p, p_0 is the probability of exposure in the control group, p_1 is the probability of exposure in the experimental group, the relative risk is p_1/p_0 and n_c is the continuity corrected sample size.

Unpaired t test

The minimum number of experimental subjects needed to detect a specified difference delta in population means is estimated for a given power and alpha (Dupont, 1990; Pearson and Hartley, 1970).

The estimated sample size n is calculated as the solution of:

$$n = \frac{(t_{n(m+1)-2,\alpha/2} + t_{n(m+1)-2,\beta})^2}{d^2}$$

- where d = delta/sd when delta is the difference in population means and sd is the estimated standard deviation for within group differences, α is alpha, β is 1 -

power, $t_{\nu,p}$ is a Student t quantile with ν degrees of freedom and probability p and m is the number of control subjects per experimental subject.

Paired t test

The minimum number of pairs of subjects needed to detect a specified difference delta in population means is estimated for a given power and alpha. (Dupont, 1990; Pearson and Hartley, 1970).

The estimated sample size n is calculated as the solution of:

$$n = \frac{(t_{n-1,\alpha/2} + t_{n-1,\beta})^2}{d^2}$$

- where d = delta/sd when delta is the difference in population means and sd is the estimated standard deviation of paired response differences; α is alpha; β is 1 - power and $t_{\nu,p}$ is a Student t quantile with ν degrees of freedom and probability p.

Survival times (two groups)

The minimum number of subjects required to detect a specified ratio of median survival times is estimated for a given power and alpha (Dupont, 1990; Schoenfeld and Richter, 1982).

The estimated sample size n is calculated as:

$$n = \left(Z_{\alpha/2} + Z_{\beta}\right)^2 2\left(\frac{2/p}{\ln(r)}\right)^2$$

$$p = 1 - p_a \exp\left(-\ln(2)\frac{F}{m}\right)$$

$$p_a = \frac{1 - \exp\left(-\ln(2)\frac{A}{m}\right)}{\ln(2)\frac{A}{m}}$$

$$m = (C + E)/2$$

- where α = alpha, β = 1 - power and Zp is the standard normal deviate for probability p.

Randomization

Random sequences and random numbers are generated by application or transformation of the output from a uniform random number generator.

Uniform random deviates are generated using the combined RANROT type W and Mother-of-All algorithm described by Fog (2000). This algorithm uses a word length of 32-bits and provides 63-bit resolution; it passes all of the DIEHARD tests (Marsaglia, 1997) and performs well in the theoretical spectral tests (Knuth, 1997).

StatsDirect seeds the random number generator with a number taken from the computer's clock (the number of hundredths of a second which have elapsed since midnight). It is highly improbable that StatsDirect will produce the same "random" sequence more than once, the time is stamped on randomization output that this can be validated. The user can specify seeds for the generation of series of random numbers.

Random number generation functions are provided for uniform, normal, binomial, Poisson, gamma and exponential distributions. The normal/Gaussian and exponential random generators use a transformation of uniform deviates whereas the gamma, Poisson and binomial random generators use rejection methods (Press et al., 1992).

Randomization is provided for experimental design. The user can use StatsDirect to randomize a series of integers, a given number of case-control pairs or a given number of subjects equally to two independent groups. A group allocation routine is also provided for the situation of m ($\leq$ k) weighted preferences for the allocation of n subjects to k groups.

Proportions (binomial)

Tests are provided for the differences between and ratios of binomial proportions (only one of two possible outcomes for each observation, numerator as count of outcomes in one direction and denominator as number of trials run).

Single

An observed single binomial proportion is compared with an expected proportion (binomial parameter). An exact confidence interval and an approximate mid-P confidence interval are provided for the proportion. Exact P and exact mid-P hypothesis tests for the equality (null hypothesis) of observed and expected proportions (Armitage and Berry, 1994; Gardner and Altman, 1989).

The Clopper-Pearson method is used for the exact confidence interval and the Newcombe-Wilson method is used for the mid-P confidence interval (Newcombe, 1998c).

Paired

Exact and exact mid-P hypothesis tests are calculated for the equality (null hypothesis) of a pair of proportions and constructs a confidence interval for the difference between them. Exact methods are used wherever practical (Armitage and Berry, 1994; Liddell, 1983).

The two sided exact P value equates with the exact test for a paired fourfold table (Liddell, 1983). With large numbers, an appropriate normal approximation is used in the hypothesis test (most asymptotic methods tend to mid-P).

The confidence interval is constructed using Newcombe's refinement of Wilson's score based method; this is close to a mid-P interval (Newcombe, 1998a).

Two independent

An hypothesis test around the difference between the two proportions is tested and a confidence interval constructed. An exact two sided P value is calculated for the hypothesis test (null hypothesis that there is no difference between the two proportions) using a mid-P approach to Fisher's exact test. The conventional normal approximation is also given for the hypothesis test and the user is informed in the help system that this is for use only with large numbers (Armitage and Berry, 1994).

The iterative method of Miettinen and Nurminen is used to construct the confidence interval for the difference between the proportions (Mee, 1984; Anbar, 1983; Gart and Nam, 1990; Miettinen and Nurminen, 1985; Newcombe, 1998a). This "near exact" confidence interval will be in close but not exact agreement with the exact two sided (mid) P value; i.e. just excluding zero and just exceeding P = 0.05.

Chi-square methods

Chi-square methods are presented for testing the association between various forms of classification. Methods for common two-dimensional contingency tables, some with stratification, are presented in an interactive form that is presented in the style of explanatory textbooks (Armitage and Berry, 1994; Agresti, 1996; Bland, 1996).

Two by two tables

The basic chi-square statistic for testing association is calculated as:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- where, for r rows and c columns of n observations, O is an observed frequency and E is an estimated expected frequency. The expected frequency for any cell is estimated as the row total times the column total then divided by the grand total (n).

Yates' continuity correction improves the approximation of the discrete sample chi-square statistic to a continuous chi-square distribution (Armitage and Berry, 1994):

$$Yates'\,corrected\ \chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(\left|O_{ij} - E_{ij}\right| - 0.5\right)^2}{E_{ij}}$$

The user is given the option to calculate Fisher's exact test and is informed in the help text of Cochrane's rules (no expected frequency should be less than 1 and at least 80% of expected frequencies should be greater than 5).

The user is prompted to specify the nature of the data; if they are from a case-control study then the odds ratio (with confidence interval) is calculated; if they are from a cohort study then the relative risk (with confidence interval) is calculated.

Two by k tables

A two by k chi-square test is calculated for testing independence and linear trend in a series of k proportions.

The basic statistic for independence is calculated as:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- where, for r rows and c columns of n observations, O is an observed frequency and E is an estimated expected frequency. The expected frequency for any cell is estimated as the row total times the column total divided by the grand total (n).

The linear trend statistic is calculated as (Armitage and Berry, 1994):

$$\chi^2_{trend} = \frac{\left[\sum_{i=1}^{k} r_i v_i - R\mu\right]^2}{p(1-p)\left[\sum_{i=1}^{k} n_i v_i^2 - N\mu^2\right]}$$

$$\mu = \sum_{i=1}^{k}\frac{n_i v_i}{N}$$

- where each of k groups of observations are denoted as r_i successes out of n_i total with score v_i assigned. R is the sum of all r_i, N is the sum of all n_i and p = R/N.

r by c tables

Two asymptotic tests of independence are performed, chi-square and G-square (likelihood-ratio chi-square). Both test statistics indicate the degree of independence between the variables that make up the table. The G-square statistic is less reliable than the chi-square statistic when numbers are small.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\log\left(\frac{O_{ij}}{E_{ij}}\right)$$

- where, for r rows and c columns of n observations, O is an observed frequency and E is an estimated expected frequency. The expected frequency for any cell is estimated as the row total times the column total divided by the grand total (n).

An exact permutational test of independence is also performed. This test is a generalisation of the Fisher-Irwin method for two by two tables. The network search algorithm described by Mehta and Patel (1983, 1986a, 1986b) is employed to find the solution without needing to enumerate all possible tables.

$$P_f = \frac{\prod_{i=1}^{r} f_{i.}!\prod_{j=1}^{c} f_{.j}!}{f_{..}!\prod_{i=1}^{r}\prod_{j=1}^{c} f_{ij}!}$$

$$P = \sum_{P\le P_0} P_f$$

- where P is the two sided Fisher probability, P_f is the conditional probability for the observed table given fixed row and column totals ($f_{i.}$ and $f_{.j}$ respectively), $f_{..}$ is the total count and ! symbolises a factorial.

Analysis of trend in r by c tables indicates how much of the general independence between scores is accounted for by linear trend. StatsDirect uses equally spaced scores for this purpose unless the user specifies otherwise.

$$R = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c} u_i v_j O_{ij} - \left(\sum_{i=1}^{r} u_i O_{i+}\right)\left(\sum_{j=1}^{c} v_j O_{j+}\right)/n}{\sqrt{\left[\sum_{i=1}^{r} u_i^2 O_{i+} - \frac{\left(\sum_{i=1}^{r} u_i O_{i+}\right)^2}{n}\right]\left[\sum_{j=1}^{c} v_j^2 O_{j+} - \frac{\left(\sum_{j=1}^{c} v_j O_{j+}\right)^2}{n}\right]}}$$

- where, for r rows and c columns of n observations, O is an observed frequency and E is an estimated expected frequency. The expected frequency for any cell is estimated as the row total times the column total divided by the grand total (n). Row scores are u, column scores are v, row totals are O_{j+} and column totals are O_{i+}.

The sample correlation coefficient R is calculated to reflect the direction and closeness of linear trend in the table. The test for linear trend is related to R by $M^2=(n-1)R^2$ which is numerically identical to Armitage's chi-square for linear trend.

The ANOVA output applies techniques similar to analysis of variance to an r by c table. Here the equality of mean column and row scores is tested. StatsDirect uses equally spaced scores for this purpose unless the user specifies scores.

Pearson's and Cramér's (V) coefficients of contingency and the phi (ϕ, correlation) coefficient reflect the strength of the association in a contingency table (Agresti, 1996; Fleiss, 1981; Stuart and Ord, 1994):

$$\phi = \sqrt{\chi^2/n}$$

$$Pearson = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

$$V = \sqrt{\frac{\chi^2/n}{\min(r-1, c-1)}}$$

McNemar

The McNemar chi-square statistic for matched pairs is calculated but the user is encouraged to use the exact equivalent and the relevant confidence intervals due to Liddell (1983) instead. See Matched Pairs in Exact Tests on Counts below.

Mantel-Haenszel

The Mantel-Haenszel chi-square test (null hypothesis of pooled odds ratio equal to zero) is presented in the context in which it is most often used, namely that of odds ratio meta-analysis. See the meta-analysis section below for further details.

Woolf

Woolf's alternative to the Mantel-Haenszel method for pooling odds ratios from several strata of fourfold tables is included mainly for educational purposes. The user is warned that the Mantel-Haenszel method is more robust, especially when some of the observed frequencies are small.

Standard weighting methods given by Armitage and Berry (1994) are used to calculate the pooled values and detailed intermediate statistics are given for each stratum.

Goodness of fit

A distribution of classes of observations is compared with an expected distribution. The user is asked to proved data that consist of a random sample of independent observations, the expected distribution of which is specified (Armitage and Berry, 1994; Conover, 1999).

Pearson's chi-square goodness of fit test statistic is calculated as:

$$T = \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j}$$

- where O_j are observed counts, E_j are corresponding expected count and c is the number of classes for which counts/frequencies are being analysed.

The user is warned that "the test has relatively low power (chance of detecting a real effect) with all but large numbers or big deviations from the null hypothesis (all classes contain observations that could have been in those classes by chance)".

The user is also warned that the handling of small expected frequencies is controversial. Koehler and Larnz (1980) assert that the chi-square approximation is adequate provided all of the following are true:

total of observed counts (N) ≥ 10

number of classes (c) ≥ 3

all expected values ≥ 0.25

Exact methods for counts

This section provides permutational probabilities and exact confidence limits for various counts and tables. The user is warned that the use of exact methods for analysis is not an adequate alternative to collecting larger numbers of data.

Sign test

The binomial distribution is used to evaluate the probability (under the null hypothesis) that an observed proportion is equal to 0.5. The user is informed that an equivalent test for expected proportions other than 0.5 is available via the single proportion function.

One and two sided cumulative probabilities are calculated under the null hypothesis. A normal approximation is used with large numbers. An exact confidence interval is constructed for the observed proportion using the Clopper-Pearson method (Conover, 1999; Vollset, 1993).

Fisher's exact test

Fisher's exact test is calculated using the null hypothesis that the two dimensions of classification in a fourfold table are equal.

Probability is computed by considering all possible tables that could give the row and column totals observed. Probability is calculated by computing areas of the hypergeometric distribution because the first cell in a fourfold table is hypergeometrically distributed (Conover, 1999; Shea, 1989; Berger, 1991):

$$P(T=a)=\frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} \quad x=0,1,...,\min(a+b,a+c)$$

-where T, the test statistic, is the expectation of a for the fourfold table below:

		Classification 1:	
		present	absent
Classification 2:	present	a	b
	absent	c	d

The binomial coefficient notation used above is expanded as follows:

$$\binom{y}{z} = \frac{y!}{z!(y-z)!}$$

- in order to avoid incomputably large factorials the logarithm of the gamma function is used to compute factorials generalised to a gamma function on a log scale (see Binomial Distribution).

Bailey's definition of two sided probability is employed (Bailey, 1977), here P values for all possible tables with P less than or equal to that for the observed table are summed. Many authors prefer to simply double the one sided P value - this is also presented (Armitage and Berry, 1994; Bland, 1996).

An expanded version of the test is provided for educational purposes, this demonstrates the cumulation of probability.

A generalisation of this method for r by c tables is calculated as described under the r by c chi-square methods above.

Exact confidence limits for two by two odds

Gart's method is used to construct exact confidence limits for the odds ratio of a fourfold table (Thomas, 1971).

The default limits are 95, 99 and 90 per cent (two sided). The user may also enter individual tail areas; e.g. one sided 95% confidence limit via a lower tail area of 0% and an upper tail area of 5%.

Matched pairs

Paired proportions have traditionally been compared using McNemar's chi-square test but an exact alternative due to Liddell (1983) is calculated here.

		Category 1:	
		outcome present	outcome absent
Category 2:	outcome present	a	b
	outcome absent	c	d

The exact test is a special case of the sign test. The b count in the table below is treated as a binomial variable from the sample b+c. Using the ratio R' (R' = b/c) as a point estimate of relative risk, a two sided probability is calculated that R' = 1 (the null hypothesis). The test statistic F=b/(c+1).

Confidence limits for R' are calculated as follows:

$$R'_L = \frac{b}{(c+1)F(0.025,2(c+1),2b)}$$

$$R'_U = \frac{(b+1)F(0.025,2(b+1),2c)}{c}$$

- where F(P,n,d) is a quantile from the F distribution with n and d degrees of freedom.

Miscellaneous methods

This section includes calculations for a series of epidemiological methods; basic concepts of Epidemiology are explained to the user via the help system (Rothman and Greenland, 1998). Statistical methods commonly used in Clinical Epidemiology and Evidence Based Health are not commonly found in statistical software, therefore, the methods are presented here in detail (Sackett et al., 1983, 1991; Sackett 1996).

Risk (prospective)

Calculations relevant to prospective studies of risk are performed as follows (Sahai and Kurshid, 1996):

		EXPOSED	UNEXPOSED
OUTCOME:	YES	a	b
	NO	c	d

Outcome rate exposed (P_e) = a/(a+c)
Outcome rate not exposed (P_u) = b/(b+d)

Relative risk (RR) = P_e/P_u
Risk difference (RD) = P_e-P_u

Estimate of population exposure (P_x) = (a+c)/(a+b+c+d)
Population attributable risk % = $100(P_x(RR-1))/(1+(P_x(RR-1)))$

The iterative approximation recommended by Gart and Nam is used to construct confidence intervals for relative risk (Gart and Nam, 1988).

The confidence interval for risk difference is constructed using an exact method (Mee, 1984; Anbar, 1983; Gart and Nam, 1990; Newcombe, 1998b).

Walter's approximate variance formula is used to construct the confidence interval for population attributable risk (Walter, 1978; Leung and Kupper, 1981).

Risk (retrospective)

Calculations relevant to retrospective studies of risk are performed as follows(Sahai and Kurshid, 1996):

		EXPOSED	UNEXPOSED
OUTCOME:	YES	a	b
	NO	c	d

Odds ratio (OR) = (a*d)/(b*c)

Estimate of population exposure (P_x) = c/(c+d)
Estimate of population attributable risk% = $100(P_x(OR-1))/(1+(P_x(OR-1)))$

A confidence interval for the odds ratio is calculated using two different methods. The logit method for large samples is given first followed by either Gart's (if n < 100,000) or Cornfield's method (Fleiss, 1979; Gardner and Altman, 1989; Thomas, 1971). If numbers are too large for Gart's method and a convergent solution can not been achieved with Cornfield's method then only the logit interval is given, otherwise Gart's or Cornfield's interval is given.

Number needed to treat

Calculations around risk in a clinical context of number needed to treat are performed as follows (Sackett, 1996; Sackett et al., 1983, 1991; Altman, 1991; Sahai and Kurshid, 1996; Laupacis et al., 1988):

		TREATED	NOT TREATED/CONTROLS
ADVERSE EVENT:	YES	a	b
	NO	c	d

p_c = proportion of subjects in control group who suffer an event
p_t = proportion of subjects in treated group who suffer an event

$p_c = b / (b + d)$
$p_t = a / (a + c)$

Relative risk reduction = $(p_c - p_t) / p_c = RR$
Absolute risk reduction = $p_c - p_t = ARR = RR * p_c$
Number needed to treat = $1 / (p_c - p_t) = 1 / ARR$

The consensus practice of rounding NNT statistics upward is adopted here (Sackett, 1996)

Confidence intervals for relative risk and relative risk reduction are calculated using the iterative approaches to ratios of binomial proportions described by Gart and Nam (Gart and Nam, 1988; Haynes and Sackett, 1993). Confidence intervals for absolute risk reduction and number needed to treat are based on the iterative method of Miettinen and Nurminen (Mee, 1984; Anbar, 1983; Gart and Nam 1990; Miettinen and Nurminen, 1985) for constructing confidence intervals for differences between independent binomial proportions.

Incidence rates

Person-time data from prospective studies of two groups with different exposures may be expressed as a difference between incidence rates and as a ratio of incidence rates.

	EXPOSURE:		
OUTCOME:	Exposed	Not Exposed	Total
Cases	a	b	m
Person-time	PT_1	PT_2	PT

Incidence Rate (exposed) = a/PT_1

Incidence Rate (not exposed) = b/PT_2

The exact Poisson and test-based methods described by Sahai and Kurshid (1996) are used to construct confidence intervals for incidence rate ratios and differences where there are two exposure classes:

$$\hat{IRD} = \frac{a}{PT_1} - \frac{b}{PT_2}$$

$$\chi^2 = \left(a - \frac{m\,PT_1}{PT} \right)^2 \Bigg/ \left(\frac{m\,PT_1\,PT_2}{PT^2} \right)$$

$$\hat{IRD}_L = \hat{IRD} - Z_{\frac{\alpha}{2}} \sqrt{\hat{IRD}^2 / \chi^2}$$

$$\hat{IRD}_U = \hat{IRD} + Z_{\frac{\alpha}{2}} \sqrt{\hat{IRD}^2 / \chi^2}$$

$$\hat{IRR} = \frac{a}{PT_1} \Bigg/ \frac{b}{PT_2}$$

$$\hat{IRR}_L = \left(\frac{PT_2}{PT_1} \right) \left(\frac{a}{b+1} \right) F_{\frac{\alpha}{2},2(b+1),2a}$$

$$\hat{IRR}_U = \left(\frac{PT_2}{PT_1} \right) \left(\frac{a+1}{b} \right) F_{\frac{\alpha}{2},2(a+1),2b}$$

- where IRD hat and IRR hat are point estimates of incidence rate difference and ratio respectively, Z is a quantile of the standard normal distribution and F is a quantile of the F distribution (denominator degrees of freedom are quoted last).

Diagnostic tests and likelihood ratios

Diagnostic test data presented in two by two format are analysed as follows (Sackett et al., 1983, 1991):

		DISEASE:	
		Present	Absent
TEST:	+	a (true +ve)	b (false +ve)
	-	c (false -ve)	d (true -ve)

Sensitivity = a/(a+c)

Specificity = d/(b+d)

+ve predictive value = a/(a+b)

-ve predictive value = d/(d+c)

Likelihood ratio of a positive test = [a/(a+c)]/[b/(b+d)]

Likelihood ratio of a negative test = [c/(a+c)]/[d/(b+d)]

For data presented in two by k format, at each of the k strata likelihood ratios are calculated as follows:

likelihood ratio j = $p(t_{k\_disease})/p(t_{k\_no\ disease})$

where $p(t_{k\_}$ is the proportion displaying the relevant test result at level k

Confidence intervals for the likelihood ratios are constructed using the iterative method for ratios of binomial proportions described by Gart and Nam (1988).

Screening test errors

Further to the diagnostic test analyses described above, an additional function is provided for the more specific context of evaluating screening tests. This is a calculation of the probability of false positive and false negative results with a test of given true and false positive rates and a given prevalence of disease (Fleiss, 1981).

$$P_{F+} = \frac{P(A \mid \overline{B})(1 - P(B))}{P(A \mid \overline{B}) + P(B)\left[P(A \mid B) - P(A \mid \overline{B})\right]}$$

$$P_{F-} = \frac{\left[1 - P(A \mid B)\right]P(B)}{1 - P(A \mid \overline{B}) - P(B)\left[P(A \mid B) - P(A \mid \overline{B})\right]}$$

- where P_{F+} is the false positive rate, P_{F-} is the false negative rate, P(A|B) is the probability of A given B, A is a positive test result, A bar is a negative test result, B is disease present and B bar is disease absent.

Standardised mortality ratios

The indirect method is used to calculate standardized mortality ratios (SMR) from the following data:

Groups, e.g. age bands for indirect age standardization
Mortality rates for each group from a reference population
Size of each group in the study population
Multiplier for mortality, e.g. 10000 if mortality entered as deaths per 10000, 1 if mortality entered as a decimal fraction

The SMR is expressed in ratio and integer (ratio × 100) formats with a confidence interval. The confidence intervals are calculated by the exact Poisson method of Owen, this gives better coverage than the frequently quoted Vandenbroucke method (Ulm, 1990; Greenland, 1990).

A test based on the null hypothesis that the number of observed and expected deaths are equal is also given. This test uses a Poisson distribution to calculate probability (Armitage and Berry, 1994, Bland, 1996; Gardner and Altman, 1989).

Kappa agreement statistics for two raters

Cohen's kappa (weighted and unweighted) and Scott's pi are calculated as measures of inter-rater agreement for two raters' categorical assessments (Fleiss, 1981; Altman, 1991; Scott, 1955). Data are accepted by direct interactive input or via the workbook. The user is warned of the various shortcomings and traps of using kappa.

Weighted kappa partly compensates for a problem with unweighted kappa, namely that it is not adjusted for the degree of disagreement. Disagreement is weighted in decreasing priority from the top left (origin) of the table. The user is offered the following weights (1 is the default):

- w(ij)=1-abs(i-j)/(g-1)
- w(ij)=1-[(i-j)/(g-1)]²
- User defined (only available via workbook data entry)

g = categories, w = weight
i = category for one observer (from 1 to g)
j = category for the other observer (from 1 to g)
In broad terms a kappa below 0.2 indicates poor agreement and a kappa above 0.8 indicates very good agreement beyond chance. The following guide is given to the user (Landis and Koch, 1977):

Kappa	Strength of agreement
< 0.2	Poor
> 0.2 ≤ 0.4	Fair
> 0.4 ≤ 0.6	Moderate
> 0.6 ≤ 0.8	Good
> 0.8 1	Very good

The user is warned that kappa values from different studies can not be reliably compared because kappa is sensitive to the prevalence of different categories. i.e. if one category is observed more commonly in one study than another then kappa may indicate a difference in inter-rater agreement which is not due to the raters.

In the case of two categories, Scott's pi is presented as the statistic of choice (Zwick, 1988; Scott, 1955), and its confidence interval is constructed by the Donner-Eliasziw (1992) method.

Maxwell's test of marginal homogeneity is given as a method of looking for differences between the raters in at least one category (Maxwell, 1970). Maxwell's generalisation of the McNemar statistic is given as a method of investigating the spread of inter-rater differences, more specifically the symmetry of these differences about the diagonal (Maxwell, 1970).

The condition of more than two raters was deemed beyond the scope of this work because of the need for the user to have expert statistical skills in order to interpret results from methods such as those described by Fleiss (1981).

Basic univariate descriptive statistics

Basic descriptive statistics are calculated to 64 bit decimal precision avoiding any of the pocket calculator formulae that led to unnecessary lack of precision (McCullough and Wilson, 1999).

Valid and missing data

For each worksheet column that you select, the number of valid data are the number of cells that can be interpreted as numbers, the remaining cells that can not be interpreted as numbers are counted as missing (e.g. empty cell, asterisk or text label). The sample size used in the calculations below is the number of valid data.

Variance, standard deviation and standard error

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i)}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$sem = \frac{s}{\sqrt{n}}$$

$$lower\,CL = \bar{x} - sem\,t_{\alpha,n-1}$$

$$upper\,CL = \bar{x} + sem\,t_{\alpha,n-1}$$

$$vc = \frac{s}{\bar{x}}$$

- where Σ is the summation for all observations (xi) in a sample, x bar is the sample (arithmetic) mean, n is the sample size, s^2 is the sample variance, s is the sample standard deviation, sem is the standard error of the sample mean, upper and lower CL are the confidence limits of the confidence interval for the mean, $t_{\alpha,n-1}$ is the ($100 \times \alpha$% two tailed quantile from the Student t distribution with n-1 degrees of freedom, and vc is the variance coefficient.

Skewness and kurtosis

$$skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 / n}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 / n\right]^{1.5}}$$

$$kurtosis = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 / n}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 / n\right]^{2}}$$

- where Σ is the summation for all observations (xi) in a sample, x bar is the sample mean and n is the sample size. Note that there are other definitions of these coefficients used by some other statistical software. StatsDirect uses the standard definitions for which critical values are published in standard statistical tables (Pearson and Hartley, 1970; Stuart and Ord, 1994).

Geometric mean

The geometric mean is a useful measure of central tendency for samples that are log-normally distributed (i.e. the logarithms of the observations are from an approximately normal distribution). The geometric mean is not calculated for samples that contain negative values.

$$gm = \exp\left[\frac{\sum_{i=1}^{n}\ln(x_i)}{n}\right]$$

- where Σ is the summation for all observations (xi) in a sample, ln is the natural (base e) logarithm, exp is the exponent (anti-logarithm for base e), gm is the sample geometric mean and n is the sample size.

Median, quartiles and range

The user is advised that for samples that are not from an approximately normal distribution, for example when data are censored to remove very large and/or very small values, the following nonparametric statistics should be used in place of the arithmetic mean, its variance and the other parametric measures above.

Median (quantile 0.5), lower quartile (25th centile, quantile 0.25) and upper quartile (75th centile, quantile 0.75) are defined generally as quantiles:

$$Q(p) = z\,u(k+1) + (1-z)\,u(k)$$

$$j = \max\left[\min\left\{p(n+1), n\right\}, 1\right]$$

$$k = fix(j)$$

$$z = j - k$$

- where p is a proportion, Q is the p^{th} quantile (e.g. median is Q(0.5)), fix is the integer part of a real number, k is the order statistic, z is the fractional part of the order statistic (0 or 0.5), u is an observation from a sample after it has been ordered from smallest to largest value and n is the sample size.

Parametric methods

This section provides various methods for which data are assumed to have been drawn from a normal distribution. A test for certain types of non-normality is also included here.

Student's t tests

For single samples, the test statistic is calculated as:

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2 / n}}$$

- where x bar is the sample mean, s^2 is the sample variance, n is the sample size, μ is the specified population mean and t is a Student t quantile with n-1 degrees of freedom.

For paired samples, the test statistic is calculated as:

$$t = \frac{\bar{d}}{\sqrt{s^2 / n}}$$

- where d bar is the mean difference, s^2 is the sample variance, n is the sample size and t is a Student t quantile with n-1 degrees of freedom.

Limits of agreement and an agreement plot are also given for paired data. The user is reminded that if the main purpose in studying a pair of samples is to see how closely the samples agree, rather than looking for evidence of difference, then limits of agreement are useful (Bland and Altman 1986, 1996a, 1996b).

For two independent samples, the test statistic is calculated as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{j=1}^{n_1}(x_j - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_i - \bar{x}_2)^2}{n_1 - n_2 - 2}$$

- where x_1 bar and x_2 bar are the sample means, s^2 is the pooled sample variance, n_1 and n_2 are the sample sizes and t is a Student t quantile with $n_1 + n_2 - 2$ degrees of freedom.

The user is warned that the unpaired t test must not be used if there is a significant difference between the variances of the two samples; StatsDirect tests for this and gives appropriate warnings.

Normal distribution tests

Normal distribution tests are provided here as large (say > 50) sample methods for parity with textbooks. Ordinarily, the user would be encouraged to use Student t tests.

For single samples (also differences between pairs), the test statistic is calculated as:

$$z = \frac{\bar{x} - \mu}{\sqrt{s^2 / n}}$$

- where x bar is the sample mean, s^2 is the sample variance, n is the sample size, μ is the specified population mean and z is a quantile from the standard normal distribution.

For two independent samples, the test statistic is calculated as:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_2^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- where x_1 bar and x_2 bar are the sample means, s_1^2 and s_2^2 are the sample variances, n_1 and n_2 are the sample sizes and z is a quantile from the standard normal distribution.

For log-normal data, geometric mean (arithmetic mean of logs) and reference range is calculated as:

$$e^{g \pm z s}$$

$$g = \sum_{i=1}^{n} \ln(x_i)$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} \ln(x_i)^2 - \left[\sum_{i=1}^{n} \ln(x_i)\right]^2 \Big/ n}{n-1}}$$

- where e is exponent, g is geometric mean, ln is natural logarithm, n is sample size and z is a quantile from the standard normal distribution (alpha/2 quantile for a 100*(1-alpha)% confidence interval).

Reference ranges

Reference ranges/intervals are constructed by normal theory (Altman, 1991):

$$rr = \bar{x} \pm z_{(1-c)/2} s$$

$$se = \sqrt{\frac{s^2}{n} + \frac{z^2_{(1-c)/2}\, s^2}{2n}}$$

$$ci = rr \pm z_{\alpha/2} se$$

- where x bar is the sample mean, s is the sample standard deviation, n is the sample size, rr is the reference range, se is the standard error of the reference range limits, ci is the confidence interval for the reference range limits, z is a quantile from the standard normal distribution and c is per cent range coverage/100 (e.g. 0.95 for a 95% reference range).

For samples with no negative values, the above calculations are repeated on log-transformed data and the results are presented in the original measurement scale.

The percentile method is also given for samples that are not from approximately normal or log-normal distributions. For a c*100% reference range, the percentile method examines the 1-(c/2) and 1-(1-(c/2)) sample quantiles and their confidence intervals as described under "quantile confidence interval" below.

Poisson confidence intervals

The Poisson mean is estimated as the arithmetic mean of the sample and the confidence interval is estimated using the relationship between the chi-square and Poisson distributions (Stuart and Ord, 1994; Johnson and Kotz, 1969):

$$LL = \frac{\chi^2_{2n,\alpha/2}}{2}$$

$$UL = \frac{\chi^2_{2n+2,1-\alpha/2}}{2}$$

- where $\chi^2_{\alpha,\nu}$ is the chi-square deviate with lower tail area α on ν degrees of freedom, n is the sample size and the sample mean is the point estimate of the

Poisson parameter around which LL and UL are the lower and upper confidence limits respectively.

Shapiro-Wilk W test

The Shapiro-Wilk procedure is a semi/non-parametric analysis of variance that detects a broad range of different types of departure from normality in random samples of 3 to 5000 data.

The user is advised not to use parametric methods with samples for which the W statistic is significant (null hypothesis is that the sample is taken from a normal distribution). They are also advised that this is not a test for normality and to seek expert statistical assistance if in doubt.

Most authors agree that this is the most reliable test for non-normality for small to medium sized samples (Conover, 1999; Shapiro and Wilk, 1965; Royston, 1982a, 1982b, 1995).

The Shapiro-Wilk test is adjusted for censored data (Royston, 1995). The user is warned that different techniques are required for truncated distributions (Verrill and Johnson, 1988).

Nonparametric methods

This section provides various rank-based hypothesis tests and descriptive functions that do not assume that data are from normal distributions. Exact permutations of probability are used wherever practicable.

Mann-Whitney

The test statistic is calculated as follows:

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{i=n_1+1}^{n_2} R_i$$

- where samples of size n_1 and n_2 are pooled and R_i are the ranks.
U can be resolved as the number of times observations in one sample precede observations in the other sample in the ranking.

The sampling distribution of U is enumerated to give exact probabilities, with or without tied data (Conover, 1999; Dineen and Blakesley, 1973; Harding, 1983; Neumann, 1988).

Confidence intervals are constructed for the difference between the means or medians by examination of all pairwise differences (Conover, 1999). The level of confidence used will be as close as is theoretically possible to the one specified by the user. The selected confidence level is approached from the conservative side.

When samples are large (either sample > 80 or both samples >30), a normal approximation is used for the hypothesis test and for the confidence interval.

Wilcoxon signed ranks test

The Wilcoxon signed ranks test statistic T^+ is calculated as the sum of the ranks of the positive, non-zero differences (D_i) between a pair of samples.

Exact permutational probability associated with the test statistic is calculated for sample sizes of less than 50. A normal approximation is used with sample sizes of 50 or more. Confidence limits are calculated by examination of all pairwise means; critical values for k (the critical location index in a sorted vector of all pairwise means) are used with sample sizes up to 30, and estimate of k by K* is used for samples with more than 30 observations (Conover, 1999; Neumann, 1988).

Spearman's rank correlation

Spearman's rank correlation coefficient (ρ) is calculated as:

$$\rho = \frac{\sum_{i=1}^{n} R(x_i)R(y_i) - n\left(\frac{n+1}{2}\right)^2}{\left(\sum_{i=1}^{n} R(x_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)^{0.5}\left(\sum_{i=1}^{n} R(y_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)^{0.5}}$$

- where R(x) and R(y) are the ranks of a pair of variables (x and y) each containing n observations.

Using this formula ρ is equivalent to Pearson's r based on ranks and average ranks. The probability associated with ρ is evaluated using a recurrence method when $n < 10$ and the Edgeworth series expansion when $n \geq 10$ (Best and Roberts, 1975). A confidence interval for rho is constructed using Fisher's Z transformation (Conover, 1999; Gardner and Altman, 1989; Hollander and Wolfe, 1999).

Kendall's rank correlation

Consider two samples, x and y, each of size n. The total number of possible pairings of x with y observations is n(n-1)/2. Now consider ordering the pairs by the x values and then by the y values. If $x_3 > y_3$ when ordered on both x and y then the third pair is concordant, otherwise the third pair is discordant. S is the difference between the number of concordant (ordered in the same way, n_c) and discordant (ordered differently, n_d) pairs.

Tau (τ) is related to S by:

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

If there are tied (same value) observations then τ_b is used:

$$\tau_b = \frac{S}{\sqrt{\left[n(n-1)/2 - \sum_{i=1}^{t} t_i(t_i-1)/2\right]\left[n(n-1)/2 - \sum_{i=1}^{u} u_i(u_i-1)/2\right]}}$$

- where t_i is the number of observations tied at a particular rank of x and u is the number tied at a rank of y.

In the presence of ties, the statistic τ_b is given as a variant of τ adjusted for ties (Kendall and Gibbons, 1990). When there are no ties $\tau_b = \tau$. An asymptotically distribution-free confidence interval is constructed for τ_b or τ using the variant of the method of Samra and Randles (1988) described by Hollander and Wolfe (1999).

The gamma coefficient is given as a measure of association that is highly resistant to tied data (Goodman and Kruskal, 1963):

$$\gamma = \frac{n_c - n_d}{n_c + n_d}$$

Tests for Kendall's test statistic being zero are calculated in exact form when there are no tied data, and in approximate form through a normalised statistic with and without a continuity correction (Kendall's score reduced by 1). In the presence of ties, the normalised statistic is calculated using the extended variance formula given by Hollander and Wolfe (1999). In the absence of ties the probability of null S (and thus τ) is evaluated using a recurrence formula when $n < 9$ and an Edgeworth series expansion when $n \geq 9$ (Best and Gipps, 1974). In the presence of ties the user is guided to make inference from the normal approximation (Kendall and Gibbons, 1990; Conover, 1999; Hollander and Wolfe, 1999).

Kruskal-Wallis test

The Kruskal-Wallis test statistic for k samples, each of size n_i is calculated as:

$$T = \frac{1}{s^2}\left(\sum_{i=1}^{k}\frac{R_i}{n_i} - N\frac{(N+1)^2}{4}\right)$$

- where N is the total number (all n_i) and R_i is the sum of the ranks (from all samples pooled) for the i^{th} sample and:

$$S^2 = \frac{1}{N-1}\left(\sum_{all} R_{ij}{}^2 - N\frac{(N+1)^2}{4}\right)$$

The test statistic is an extension of the Mann-Whitney test. In the presence of tied ranks the test statistic is given in adjusted and unadjusted forms, (opinion varies concerning the handling of ties). The test statistic follows approximately a chi-square distribution with k-1 degrees of freedom. P values derived from the chi-square approximation are highly satisfactory in most cases (Conover, 1999).

If the test is significant then the user is able to make multiple comparisons between the samples. The Dwass-Steel-Critchlow-Fligner (Critchlow and Fligner, 1991; Hollander and Wolfe, 1999) and Conover-Inman (Conover, 1999) methods are used to make all possible pairwise comparisons between groups.

By the Dwass-Steel-Critchlow-Fligner procedure, a contrast is considered significant if the following inequality is satisfied:

$$W_{ij} - \frac{n_i(n_i+n_j+1)}{2} \Bigg/ \sqrt{\frac{n_i n_j}{24}\left[n_i+n_j+1-\frac{\sum_{b=1}^{g_{ij}}(t_b-1)t_b(t_b+1)}{(n_i+n_j)(n_i+n_j-1)}\right]} > q_{\alpha,k},\ for 1 \le i \le j \le k$$

- where q is a quantile from the normal range distribution for k groups, n_i is size of the ith group, n_j is the size of the jth group, t_b is the number of ties at rank b and W_{ij} is the sum of the ranks for the ith group where observations for both groups have been ranked together.

The Conover-Inman procedure is simply Fisher's least significant difference method performed on ranks. A contrast is considered significant if the following inequality is satisfied:

$$\left|\frac{R_j}{n_j} - \frac{R_i}{n_i}\right| > t_{1-\alpha/2}\sqrt{S^2\,\frac{N-1-T}{N-k}}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

- where t is a quantile from the Student t distribution on N-k degrees of freedom.

Friedman test

The Iman Davenport T_2 variant of the Friedman test statistic is calculated as:

$$T_2 = \frac{(b-1)T_1}{b(k-1)-T_1}$$

- where there are k treatments and b blocks and T_1 is:

$$T_1 = \frac{(k-1)\left[\sum_{j=1}^{k} R_j{}^2 - bC_1\right]}{A_1 - C_1}$$

- where R_j is the sum of the ranks (from pooled observations) for all blocks in a one treatment and A_1 and C_1 are:

$$A_1 = \sum_{i=1}^{b}\sum_{j=1}^{k}\left[R(X_{ij})\right]^2$$

$$C_1 = bk(k+1)^2/4$$

When the test is significant, the user is enabled to make multiple comparisons between the individual samples. These comparisons are performed automatically for all possible contrasts. A contrast is considered significant if the following inequality is satisfied:

$$\left|R_j - R_i\right| > t_{1-\alpha/2}\sqrt{\frac{2(bA_1 - \sum_{j=1}^{k} R_j{}^2}{(b-1)(k-1)}}$$

- where t is a quantile from the Student t distribution on (b-1)(k-1) degrees of freedom.

The overall test statistic T_2 is calculated as above (Iman and Davenport, 1980). T_2 is approximately distributed as an F random variable with k-1 numerator and (b-1)(k-1) denominator degrees of freedom; the P value is derived in this way from the F distribution (Conover, 1999).

Cuzick's test for trend

The test statistic is calculated as follows:

$$T = \sum_{i=1}^{k} l_i R_i$$

$$L = \sum_{i=1}^{k} l_i n_i$$

$$E(T) = \frac{L(N+1)}{2}$$

$$\text{var}(T) = \frac{N+1}{12}\left(N\sum_{i=1}^{k} l_i^2 n_i - L^2 \right)$$

$$z = \frac{T - E(T)}{\sqrt{\text{var}(T)}}$$

- where R_i is the sum of the pooled ranks for the i^{th} group, l_i is the sum of scores for the i^{th} group, n_i is the sample size for the i^{th} group and N is the total number of observations. For the null hypothesis of no trend across the groups T will have mean E(T), variance var(T) and the null hypothesis is tested using the normalised test statistic z.

Values for the scores must correspond to the order that is being tested for across the groups.

A logistic distribution is assumed for errors. Probabilities for z are derived from the standard normal distribution (Cuzick, 1985).

Quantile confidence interval

Quantiles are calculated as:

$$Q(p) = z\,u(k+1) + (1-z)\,u(k)$$

$$j = \max[\min\{p(n+1), n\}, 1]$$

$$k = fix(j)$$

$$z = j - k$$

- where p is a proportion, Q is the p^{th} quantile (e.g. median is Q(0.5)), fix is the integer part of a real number, k is the order statistic, z is the fractional part of the order statistic (0 or 0.5), u is an observation from a sample after it has been ordered from smallest to largest value and n is the sample size.

The confidence interval for the quantile is taken as the k^{th} highest and the k^{th} lowest value in the sample, where k is the critical order statistic derived from the binomial distribution relevant to the quantile, or by normal approximation when the sample size is > 50 (Conover, 1999). For a c*100% confidence interval the binomial quantiles closest to a cumulative probability of (1-c)/2 and 1-(1-c)/2 are used.

Smirnov two sample test

The test statistic for the two sided two sample Smirnov test is the largest vertical distance between the empirical distribution functions. The test statistics for the one sided tests are the largest vertical distance of one empirical distribution function above the other and vice versa.

The two samples are first sorted (in the same order) and each point separation is evaluated in order to calculate the above test statistics.

P values for the test statistics are calculated by permutation of the exact distribution where practicable or by iterative approximations otherwise (Conover, 1999; Nikiforov, 1994; Kim and Jennrich 1973).

Homogeneity of variance

The squared ranks test for two samples is calculated as:

$$T = \sum_{i=1}^{n} R[U_i]^2$$

or

$$T_{1`} = \frac{T - n\overline{R^2}}{\sqrt{\frac{nm}{N(N-1)}\sum_{i=1}^{N} R_i^4 - \frac{nm}{N-1}(\overline{R^2})^2}}$$

- where U is a transform of the first sample, R are ranks from both samples pooled, T is the test statistic for use without ties, T_1 is the test statistic for use when there are tied data, n is the first sample size, m is the second sample size, N is the total number of observations, R^2 bar is the mean squared rank from both samples pooled and R^4 is a rank raised to the fourth power. All observations are transformed to their absolute deviation from the sample mean before ranking.

The squared ranks test statistic for more than two samples is calculated as:

$$T_2 = \frac{1}{D^2}\left[\sum_{i=1}^{k}\frac{S_j^2}{n_j} - N(\overline{S})^2\right]$$

$$D^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} R_j^4 - N(\overline{S})^2\right]$$

- where N is the total number of observations, n_j is the number of observations in the j^{th} sample, S^2_j is the sum of the squared ranks in the j^{th} sample, S bar is the mean of all squared ranks and R^4 is a rank raised to the fourth power. All observations are transformed to their absolute deviation from the sample mean before ranking.

Probabilities are calculated using the asymptotic approximations described by Conover (1999).

Analysis of variance

This section contains various methods for the comparison of the means of two or more samples. Intermediate calculations require estimates of variance and thus the methods are grouped under the title analysis of variance (ANOVA).

Calculation routines are given for common experimental designs and some more unusual designs found in medical research. The user is informed, via the help system, that there are many more designs for ANOVA and that complex designs are best used only under expert statistical guidance.

The relationship between ANOVA and generalised regression is explored in the help system.

One way and homogeneity

One way analysis of variance can be considered a generalisation of the two sample t test. The F statistic compares the variability between the groups to the variability within the groups:

$$F = \frac{MST}{MSE}$$

$$MST = \frac{\sum_{i=1}^{k} (T_i^2 / n_i) - G^2 / n}{k - 1}$$

$$MSE = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i-1}^{k} (T_i^2 / n_i)}{n - k}$$

- where F is the variance ratio for the overall test, MST is the mean square due to treatments/groups (between groups), MSE is the mean square due to error (within groups, residual mean square), Y_{ij} is an observation, T_i is a group total, G is the grand total of all observations, n_i is the number in group i and n is the total number of observations (Armitage and Berry, 1994; Kleinbaum et al., 1998).

The user is given the option to investigate homogeneity of variance. Both Levene and Bartlett statistics are used and the user is encouraged to make inference from the Levene method. The W50 definition of Levene test statistic (Brown and Forsythe, 1974) is used; this is essentially a one way analysis of variance on the absolute (unsigned) values of the deviations of observations from their group medians.

Multiple comparisons

A range of functions is provided for multiple comparisons (simultaneous inference), specifically all pairwise comparisons and all comparisons with a control.

Tukey (Tukey-Kramer if unequal group sizes), Scheffé, Bonferroni and Newman-Keuls methods are provided for all pairwise comparisons (Armitage and Berry, 1994; Wallenstein, 1980; Miller, 1981; Hsu, 1996; Kleinbaum et al., 1998). Dunnett's method is used for multiple comparisons with a control group (Hsu, 1996).

The user is warned about the dangers of "data dredging" and extensive guidance is built into the help system.

Two way randomized block

The F tests for two way ANOVA are the same if either or both block and treatment factors are considered fixed or random:

$$F_{between\ treatments} = \frac{MST}{MSE}$$

$$F_{between\ blocks} = \frac{MSB}{MSE}$$

$$MST = \frac{b\sum_{i=1}^{k}(\overline{Y}_{i.} - \overline{Y}_{..})^2}{k-1}$$

$$MSB = \frac{k\sum_{i=1}^{b}(\overline{Y}_{.j} - \overline{Y}_{..})^2}{k-1}$$

$$MSE = \frac{\sum_{i=1}^{k}\sum_{j=1}^{b}(Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} - \overline{Y}_{..})^2}{(k-1)(b-1)}$$

- where F is the variance ratio for tests of equality of treatment and block means, MST is the mean square due to treatments/groups (between groups), MSB is the mean square due to blocks (between blocks), MSE is the mean square due to error (within groups, residual mean square), Y_{ij} is an observation, $Y_{i.}$ bar is a treatment group mean, $Y_{.j}$ bar .is a block mean and Y bar is the grand mean of all observations (Armitage and Berry, 1994; Kleinbaum et al., 1998).

This procedure is also extended to a randomized block design with repeated observations for each treatment/block cell. Tests in the presence of repeated observations cover differences between treatment means, between block means and block/treatment interaction (Armitage and Berry, 1994; Kleinbaum et al., 1998).

Fully nested random (hierarchical)

ANOVA for a three factor fully random nested (split-plot) model is calculated as follows (Snedecor and Cochran, 1989):

$$CF = \frac{\left(\sum_{i=1}^{g} \sum_{j=1}^{s_i} \sum_{k=1}^{n_{ij}} X_{ijk} \right)^2}{N}$$

$$SS_{total} = \sum_{i=1}^{g} \sum_{j=1}^{s_i} \sum_{k=1}^{n_{ij}} X_{ijk} - CF$$

$$SS_{groups} = \sum_{i=1}^{g} \frac{\left(\sum_{j=1}^{s_i} \sum_{k=1}^{n_{ij}} X_{ijk} \right)^2}{s_i n_{ij}} - CF$$

$$SS_{subgroups\,(group\,i)} = \sum_{i=1}^{s_i} \frac{\left(\sum_{k=1}^{n_{ij}} X_{ijk} \right)^2}{n_{ij}} - \frac{\left(\sum_{j=1}^{s_i} \sum_{k=1}^{n_{ij}} X_{ijk} \right)^2}{s_i n_{ij}}$$

- where X_{ijk} is the k^{th} observation from the j^{th} subgroup of the i^{th} group, g is the number of groups, SS_{total} is the total sum of squares, SS_{groups} is the sum of squares due to the group factor, $SS_{subgroups\ (group\ i)}$ is the sum of squares due to the subgroup factor of group i, s_i is the number of subgroups in the i^{th} group, n_{ij} is the number of observations in the j^{th} subgroup of the i^{th} group and N is the total number of observations.

Hocking (1985) describes potential instabilities of this calculation; the user is therefore asked to seek expert statistical guidance before using it.

Latin square

The Latin square ANOVA for three factors without interaction is calculated as follows (Armitage and Berry, 1994; Cochran and Cox, 1957):

$$SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{a}\sum_{k=1}^{a} X_{ijk}^2 / a - G^2 / a^2$$

$$SS_{rows} = \sum_{i=1}^{a} R_i^2 / a - G^2 / a^2$$

$$SS_{columns} = \sum_{j=1}^{a} C_j^2 / a - G^2 / a^2$$

$$SS_{treatments} = \sum_{k=1}^{a} T_k^2 / a - G^2 / a^2$$

- where X_{ijk} is the observation from the i^{th} row of the j^{th} column with the k^{th} treatment, G is the grand total of all observations, R_i is the total for the i^{th} row, C_j is the total for the j^{th} column, T_k is the total for the k^{th} treatment, SS_{total} is the total sum of squares, SS_{rows} is the sum of squares due to the rows, $SS_{columns}$ is the sum of squares due to the columns, $SS_{treatments}$ is the sum of squares due to the treatments and a is the number of rows, columns or treatments.

Crossover

Statistics for the analysis of crossover trials, with optional baseline run-in observations, are calculated as follows (Armitage and Berry, 1994; Senn, 1993):

$$ds = \sum_{i=1}^{m}\left[X_{Di} - X_{Pi}\right] + \sum_{j=1}^{n}\left[X_{Dj} - X_{Pj}\right]$$

$$dss = \sum_{i=1}^{m}\left[\left(X_{Di} - X_{Pi}\right)^2\right] + \sum_{j=1}^{n}\left[\left(X_{Dj} - X_{Pj}\right)^2\right]$$

$$t_{relative} = \frac{ds/(m+n)}{\sqrt{\left[dss - \frac{ds^2}{n+m}\right]/(m+n-1)}}$$

$$dsa = \sum_{i=1}^{m}\left[X_{Di} - X_{Pi}\right]$$

$$dsb = \sum_{j=1}^{n}\left[X_{Dj} - X_{Pj}\right]$$

$$dssa = \sum_{i=1}^{m}\left[\left(X_{Di} - X_{Pi}\right)^2\right]$$

$$dssb = \sum_{j=1}^{n}\left[\left(X_{Dj} - X_{Pj}\right)^2\right]$$

$$v = \frac{\left[dssa - \frac{dsa^2}{m}\right] + \left[dssb - \frac{dsb^2}{n}\right]}{n+m-2}$$

$$t_{treatment} = \frac{ssa/m - ssb/n}{\sqrt{v(1/m + 1/n)}}$$

$$t_{period} = \frac{ssa/m + ssb/n}{\sqrt{v(1/m + 1/n)}}$$

$$ssa = \sum_{i=1}^{m} \left[X_{Di} + X_{Pi} \right]$$

$$ssb = \sum_{j=1}^{n} \left[X_{Dj} + X_{Pj} \right]$$

$$sssa = \sum_{i=1}^{m} \left[\left(X_{Di} + X_{Pi} \right)^2 \right]$$

$$sssb = \sum_{j=1}^{n} \left[\left(X_{Dj} + X_{Pj} \right)^2 \right]$$

$$v = \frac{\left[sssa - \frac{ssa^2}{m} \right] + \left[sssb - \frac{ssb^2}{n} \right]}{n + m - 2}$$

$$t_{tp} = \frac{ssa/m - ssb/n}{\sqrt{v(1/m + 1/n)}}$$

- where m is the number of observations in the first group (say drug first); n is the number of observations in the second group (say placebo first); X_{Di} is an observation from the drug treated arm in the first group; X_{Pi} is an observation from the placebo arm in the first group; X_{Dj} is an observation from the drug treated arm in the second group; X_{Pj} is an observation from the placebo arm in the second group; $t_{relative}$ is the test statistic, distributed as Student t on $n+m-1$ degrees of freedom, for the relative effectiveness of drug vs. placebo; t_{tp} is the test statistic, distributed as Student t on $n+m-2$ degrees of freedom, for the treatment-period interaction; and $t_{treatment}$ and t_{period} are the test statistics, distributed as Student t on $n+m-2$ degrees of freedom, for the treatment and period effect sizes respectively (null hypothesis = 0). Any baseline observations are subtracted from the relevant observations before the above are calculated.

Agreement

The function calculates a one way random effects intra-class correlation coefficient, estimated within-subjects standard deviation and a repeatability coefficient (Bland and Altman 1996a and 1996b, McGraw and Wong, 1996).

Intra-class correlation coefficient is calculated as:

$$r_I = \frac{mSS_{subjects} - SS_{total}}{(m-1)SS_{total}}$$

- where m is the number of observations per subject, $SS_{subjects}$ is the sum of squared between subjects and SS_{total} is the total sum of squares (as per one way ANOVA above).

Within-subjects standard deviation is estimated as the square root of the residual mean square from one way ANOVA.

The repeatability coefficient is calculated as:

$$C_r = \sqrt{m}\, Z\, \zeta_w$$

- where m is the number of observations per subject, Z is a quantile from the standard normal distribution (usually taken as the 5% two tailed quantile of 1.96) and ζ_w is the estimated within-subjects standard deviation (calculated as above).

Intra-subject standard deviation is plotted against intra-subject means and Kendall's rank correlation is used to assess the interdependence of these two variables.

An agreement plot is constructed by plotting the maximum differences from each possible intra-subject contrast against intra-subject means and the overall mean is marked as a line on this plot.

A Q-Q plot is given; here the sum of the difference between intra-subject observations and their means are ordered and plotted against an equal order of chi-square quantiles.

The user is warned that analysis of agreement is a complex matter best carried out only under expert statistical guidance.

Regression and correlation

This section (overlapping with analysis of variance, survival analysis and nonparametric) contains a range of methods that use regression and/or correlation. The scope for misconception around these techniques is addressed in the help system that covers basic principles plus introductions to more advanced concepts and pointers to further information.

Simple linear

Regression parameters for a straight line model ($Y = a + bx$) are calculated by the least squares method (minimisation of the sum of squares of deviations from a straight line). After differentiating, the following formulae are obtained for the slope (b) and the Y intercept (a) of the line:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a = \bar{Y} - b\bar{x}$$

Pearson's product moment correlation coefficient (r) is given as a measure of linear association between the two variables:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Confidence limits are constructed for r using Fisher's z transformation. The null hypothesis that $r = 0$ (i.e. no association) is evaluated using a modified t test (Armitage and Berry, 1994; Altman, 1991).

Multiple (general) linear

The term multiple regression is applied to linear prediction of one outcome from several predictors. The general form of linear regression is given as:

$Y' = b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k$

- where Y' is the predicted outcome value for the linear model with regression coefficients $b_{1\ to\ k}$ and Y intercept b_0 when the values for the predictor variables are $x_{1\ to\ k}$.

QR decomposition by Givens rotations is used to solve the linear equations to a high level of accuracy (Gentleman, 1974; Golub and Van Loan, 1983). Predictors that are highly correlated with other predictors are dropped from the model (users warned of this in the results). If the QR method fails (rare) then singular value decomposition is used to solve the system (Chan, 1982).

The user is asked to examine residuals to assess the suitability of the model and to identify influential data. Standard error for the predicted Y, leverage h_i (the i^{th} diagonal element of the hat (XXi) matrix), Studentized residuals, jackknife residuals, Cook's distance and DFIT are also calculated (Belsley et al., 1980; Kleinbaum et al., 1998; Draper and Smith, 1998).

$$e_i = Y_i - \hat{Y}_i$$

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$$

$$r_i = \frac{e_i}{s\sqrt{1-h_i}}$$

$$s_{-i}^2 = \frac{(n-p)s^2 - e_i^2/(1-h_i)}{n-p-1}$$

$$r_{-i} = \frac{e_i}{\sqrt{s_{-i}^2(1-h_i)}}$$

$$d_i = \left\{\frac{e_i}{s\sqrt{1-h_i}}\right\}^2 \left\{\frac{h_i}{1-h_i}\right\}\frac{1}{p}$$

$$DFIT_i = \sqrt{d_i p \frac{s^2}{s^2(i)}}$$

- where p is the number of parameters in the model, n is the number of observations, e_i is a residual, r_i is a Studentized residual, r_{-i} is a jackknife residual, s^2 is the residual mean square, s^2_{-i} is an estimate of s^2 after deletion of the i^{th} residual, h_i is the leverage (i^{th} diagonal element of the hat or XXi matrix), d_i is Cook's distance and $DFIT_i$ is DFFITS (a variant of Cook's distance presented by Belsley et al. (1980)).

The multiple correlation coefficient (R) is given as Pearson's product moment correlation between the predicted values and the observed values (Y' and Y). Just as r^2 is the proportion of the total variance (s^2) of Y that can be explained by the linear regression of Y on X, R^2 is the proportion of the variance explained by the multiple regression. The significance of R is tested by the F statistic of the analysis of variance for the regression.

An adjusted value of R^2 is given as R_a^2:

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{p-1}$$

The adjustment allows comparison of R_a^2 between different regression models by compensating for the fact that R^2 is bound to increase with the number of predictors in the model. The Durbin Watson test statistic is given as a test for certain types of serial correlation (Draper and Smith, 1998).

Automatic selection of predictors is handled by calculating all possible linear regression models with different combinations of the predictors given and assessed for minimum Mallow's Cp statistic or maximum overall variance ratio (Draper and Smith, 1998). The user is advised to place more emphasis on their "real world" selection of predictors than upon purely numerical optimisation procedures such as these.

Grouped linear and test for linearity

Linearity is tested by analysis of variance for the linear regression of k outcome observations for each level of the predictor variable (Armitage, 1994):

$$SS_{regression} = \frac{\left[\sum_{i=1}^{N} x_i Y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} Y_i\right)/N\right]^2}{\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 / N}$$

$$SS_{repeats} = \sum_{i=1}^{N} Y_i^2 - \sum_{j=1}^{k} (T_j^2 / n_j)$$

$$SS_{total} = \sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} Y_i\right)^2 / N$$

- where $SS_{regression}$ is the sum of squares due to the regression of Y on x, $SS_{repeats}$ is the part of the usual residual sum of squares that is due to variation within repeats of the outcome observations, SS_{total} is the total sum of squares and the remainder represents the sum of squares due to deviation of the means of repeated outcome observations from the regression. Y is the outcome variable, x is the predictor variable, N is the total number of Y observations and n_j is the number of Y repeats for the j^{th} x observation.

Slopes of several regression lines are compared by analysis of variance as follows (Armitage, 1994):

$$SS_{common} = \frac{\left(\sum_{j=1}^{k} SxY_j\right)^2}{\sum_{j=1}^{k} Sxx_j}$$

$$SS_{between} = \sum_{j=1}^{k} \frac{(SxY_j)^2}{Sxx_j} - \frac{\left(\sum_{j=1}^{k} SxY_j\right)^2}{\sum_{j=1}^{k} Sxx_j}$$

$$SS_{total} = \sum_{j=1}^{k} SYY_j$$

- where SS_{common} is the sum of squares due to the common slope of k regression lines, $SS_{between}$ is the sum of squares due to differences between the slopes, SS_{total} is

the total sum of squares and the residual sum of squares is the difference between SS_{total} and SS_{common}. Sxx_j is the sum of squares about the mean x observation in the j^{th} group, SxY_j is the sum of products of the deviations of xY pairs from their means in the j^{th} group and SYY_j is the sum of squares about the mean Y observation in the j^{th} group.

Vertical separation of slopes of several regression lines is tested by analysis of covariance as follows (Armitage, 1994):

$$SxY_{between} = \left[\sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n_k} x_i \right)\left(\sum_{i=1}^{n_k} Y_i \right)}{n_k} \right] - \frac{\left(\sum_{i=1}^{N} x_i \right)\left(\sum_{i=1}^{N} Y_i \right)}{N}$$

$$SxY_{within} = \left[\sum_{i=1}^{N} x_i Y_i \right] - \left[\sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n_k} x_i \right)\left(\sum_{i=1}^{n_k} Y_i \right)}{n_k} \right]$$

$$SxY_{total} = \left[\sum_{i=1}^{N} x_i Y_i \right] - \frac{\left(\sum_{i=1}^{N} x_i \right)\left(\sum_{i=1}^{N} Y_i \right)}{N}$$

$$SS_{within} = SYY_{within} - \frac{(SxY_{within})^2}{Sxx_{within}}$$

$$SS_{total} = SYY_{total} - \frac{(SxY_{total})^2}{Sxx_{total}}$$

- where SS are corrected sums of squares within the groups, total and between the groups (subtract within from total). The constituent sums of products or squares are partitioned between groups, within groups and total as above.

Polynomial

A k order/degree polynomial is fitted to the user's data:

$$\hat{Y} = b_0 + b_1 x + b_2 x^2 + \ldots + b_k x^k$$

- where *Y caret* is the predicted outcome value for the polynomial model with regression coefficients b_1 *to k* for each degree and *Y* intercept b_0. The function is fitted via a general linear regression model with *k* predictors raised to the power of *i* where *i=1 to k* (Kleinbaum et al., 1998; Armitage and Berry, 1994). QR decomposition by Givens rotations is used to solve the linear equations to a high level of accuracy (Gentleman, 1974; Golub and Van Loan, 1983). If the QR method fails (rare) then singular value decomposition is used to solve the system (Chan, 1982).

The user is given an option to calculate the area under the fitted curve by two different methods. The first method integrates the fitted polynomial function from the lowest to the highest observed predictor value using Romberg's integration. The second method uses the trapezoidal rule directly on the data to provide a crude estimate of area.

Linearized estimates

Three linearizing transformations are applied to raw data and a linear model is fitted to the results by simple linear regression. The transformations are:

Exponential

Data are linearized by logarithmic transformation of the predictor (x) variable. Simple linear regression of Y vs. ln(x) gives a = *ln(intercept)* and b = *slope* for the function:

$$Y = a + e^{bx}$$

Geometric

Data are linearized by logarithmic transformation of both variables. Simple linear regression of ln(Y) vs. ln(x) gives a = *ln(intercept)* and b = *slope* for the function:

$$Y = a + x^{b}$$

Hyperbolic

Data are linearized by reciprocal transformation of both variables. Simple linear regression of 1/Y vs. 1/x gives a = *slope* and b = *intercept* for the function:

$$Y = \frac{x}{a + bx}$$

The user is warned that the errors of the outcome/response variable might not be from a normal distribution. The help system explains the shortcomings of forcing inappropriate regression models to data.

Probit analysis

Probit or logit sigmoid curves are fitted for stimulus-quantal response data in the context of classical probit analysis (Finney, 1971, 1978).

$$Y' = \Phi^{-1}(p)$$

- where Y' is the probit transformed value (5 used to be added to avoid negative values in hand calculation), p is the proportion (p = responders/total number) and inverse Φ (p) is the 100*p% quantile from the standard normal distribution.

The curve is fitted by maximum likelihood estimation with Newton-Raphson iteration. A dummy variable is used to factor in the background/natural response rate if the user specifies a response in controls.

Confidence intervals are calculated for stimulus levels associated with given response quantiles, for example ED_{50} (effective dose for 50% response) is calculated by interpolation of the stimulus that gives median response in the fitted sigmoid model.

Logistic regression

A logistic model is fitted and analysed for a binary outcome/response and one or more predictors:

$\text{logit}(Y)' = b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k$

- where *logit(Y)'* is the predicted logit transform of the outcome value for the linear model with regression coefficients $b_{1\ to\ k}$ and *Y* intercept b_0 when the values for the predictor variables are $x_{1\ to\ k}$.

$\text{logit} = \text{log odds} = \log(\pi/(1-\pi))$

- where π is the proportional response, i.e. *r* out of *n* responded so $\pi = r/n$

The logits of the response data are fitted using an iteratively re-weighted least squares method to find maximum likelihood estimates of the model parameters (McCullagh and Nelder, 1989; Cox and Snell, 1989; Pregibon, 1981).

Deviance is given as minus twice the log likelihood ratio for models fitted by maximum likelihood (Hosmer and Lemeshow, 1989; Cox and Snell, 1989; Pregibon, 1981). The value of adding parameter to a logistic model is tested by subtracting the deviance of the model with the new parameter from the deviance of the model without the new parameter, this difference is then tested against a chi-square distribution with degrees of freedom equal to the difference between the degrees of freedom of the old and new models. A model analysis option enables the user to test the model they specify against a model with only one parameter, the intercept; this tests the combined value of the specified predictors/covariates in the model.

Residuals and case-wise diagnostic statistics are calculated as follows (Hosmer and Lemeshow, 1989; Pregibon, 1981):

Leverages are the diagonal elements of the logistic equivalent of the hat matrix in general linear regression (where leverages are proportional to the distances of the j^{th} covariate pattern from the mean of the data). The j^{th} diagonal element of the logistic equivalent of the hat matrix is calculated as:

$$h_j = m_j\hat{\pi}(\mathbf{x}_j)[1-\hat{\pi}(\mathbf{x}_j)](1,\mathbf{x'}_j)(\mathbf{X'VX})^{-1}(1,\mathbf{x'}_j)'$$

- where m_j is the number of trials with the j^{th} covariate pattern, π hat is the expected proportional response, x_j is the j^{th} covariate pattern, X is the design matrix containing all covariates (first column as 1 if intercept calculated) and V is a matrix with the general element π hat(1- π hat).

Deviance residuals are used to detect ill-fitting covariate patterns, and they are calculated as:

$$d_j = \pm\sqrt{2\left[y_j \ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j)\ln\left(\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right)\right]}$$

sign for d_j is same as $(y_j - m_j\hat{\pi}_j)$

if $y_j = 0$ then $d_j = -\sqrt{2m_j\left|\ln(1-\hat{\pi}_j)\right|}$

if $y_j = m_j$ then $d_j = \sqrt{2m_j\left|\ln(\hat{\pi}_j)\right|}$

- where m_j is the number of trials with the j^{th} covariate pattern, π hat is the expected proportional response and y_j is the number of events with the j^{th} covariate pattern.

Pearson residuals are used to detect ill-fitting covariate patterns, and they are calculated as:

$$r_j = \frac{y_j - m_j\hat{\pi}}{\sqrt{m_j\hat{\pi}(1-\hat{\pi})}}$$

- where m_j is the number of trials with the j^{th} covariate pattern, π hat is the expected proportional response and y_j is the number of events with the j^{th} covariate pattern.

Standardized Pearson residuals are used to detect ill-fitting covariate patterns, and they are calculated as:

$$rs_j = \frac{r_j}{\sqrt{1-h_j}}$$

- where r_j is the Pearson residual for the j^{th} covariate pattern and h_j is the leverage for the j^{th} covariate pattern.

Deletion displacement (delta beta) measures the change caused by deleting all observations with the j^{th} covariate pattern. The statistic is used to detect observations that have a strong influence upon the regression estimates. This change in regression coefficients is calculated as:

$$\Delta\beta_j = \frac{r_j^2 h_j}{1-h_j}$$

- where r_j is the Pearson residual for the j^{th} covariate pattern and h_j is the leverage for the j^{th} covariate pattern.

Standardized deletion displacement (std delta beta) measures the change caused by deleting all observations with the j^{th} covariate pattern. The statistic is used to detect observations that have a strong influence upon the regression estimates. This change in regression coefficients is calculated as:

$$\Delta\beta s_j = \frac{rs_j^2 h_j}{1-h_j}$$

- where rs_j is the standardized Pearson residual for the j^{th} covariate pattern and h_j is the leverage for the j^{th} covariate pattern.

Deletion chi-square (delta chi-square) measures the change in the Pearson chi-square statistic (for the fit of the regression) caused by deleting all observations with the j^{th} covariate pattern. The statistic is used to detect ill-fitting covariate patterns. This change in chi-square is calculated as:

$$\Delta\chi_j^2 = \frac{r_j^2}{1-h_j}$$

- where r_j is the Pearson residual for the j^{th} covariate pattern and h_j is the leverage for the j^{th} covariate pattern.

Differences between two fitted logits have the following useful properties:

$$\text{logit}_a - \text{logit}_b = \log\left(\frac{\hat{\pi}_a}{1-\hat{\pi}_a}\right) - \log\left(\frac{\hat{\pi}_b}{1-\hat{\pi}_b}\right)$$

$$\text{logit}_a - \text{logit}_b = \log\left(\frac{\hat{\pi}_a}{1-\hat{\pi}_a} \Big/ \frac{\hat{\pi}_b}{1-\hat{\pi}_b}\right)$$

$$\text{logit}_a - \text{logit}_b = \text{odds ratio}$$

Thus the exponents of the coefficients for predictors in a fitted logistic model represent the odds ratios associated with those predictors, approximate confidence intervals are given for these odds ratios by exponentiating the confidence intervals for the parameter estimates.

A receiver operating characteristic curve is constructed by varying the proportional response between 0 and 1 in the fitted model.

Principal components

Singular value decomposition (SVD) is used to calculate the variance contribution of each component of a correlation or covariance matrix (Krzanowski, 1988; Chan, 1982):

The SVD of an *n* by *m* matrix **X** is **UΣV' = X**. **U** and **V** are orthogonal matrices, i.e. **V' V** = **V V'** where **V'** is the transpose of **V**. **U** is a matrix formed from column vectors (m elements each) and **V** is a matrix formed from row vectors (n elements each). **Σ** is a symmetrical matrix with positive diagonal entries in non-increasing order. If **X** is a mean-centred, *n* by *m* matrix where $n>m$ and rank $r = m$ (i.e. full rank) then the first *r* columns of **V** are the first *r* principal components of **X**. The positive eigenvalues of **X'X** on **XX** are the squares of the diagonals in **Σ**. The coefficients or latent vectors are contained in **V**.

Principal component scores are derived from **U** and **Σ** as **trace{(X-Y)(X-Y)'}**. For a correlation matrix, the principal component score is calculated for the standardized variable, i.e. the original datum minus the mean of the variable then divided by its standard deviation.

Cronbach's alpha is calculated to help the user to assess the effect of dropping variables from the model, for example dropping a question from a questionnaire. The overall alpha and the alpha that would be obtained if each variable in turn was dropped are calculated. If the deletion of a variable causes an increase in alpha of more than 1.0 then the user is advised to drop the variable, however, the user is also warned to reflect upon the "real world" relevance of that variable (Streiner and Norman, 1989; McDowell and Newell, 1987; Cronbach, 1951).

Survival analysis

This section provides a range of methods for the analysis of time to event, failure time or survival data. Right-censored data are factored into calculations and identified by the common coding system of 0 for uncensored and 1 for censored observations.

Kaplan-Meier

The survival rate is expressed as the survivor function (S):

$$S(t) = \frac{number\ of\ individuals\ surviving\ longer\ than\ t}{total\ number\ of\ individuals\ studied}$$

- where t is a time period known as the survival time, time to failure or time to event (such as death). *S* is also presented as the estimated probability of surviving to time *t* for those alive just before *t* multiplied by the proportion of subjects surviving to *t*.

The product limit (PL) method of Kaplan and Meier (1958) is used to estimate *S*:

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i} \right)$$

- where t_i is duration of study at point i, d_i is number of deaths up to point i and n_i is number of individuals at risk just prior to t_i. *S* hat is based upon the probability that an individual survives at the end of a time interval, on the condition that the individual was present at the start of the time interval. *S* hat is the product (Π) of these conditional probabilities.

If a subject is last followed up at time t_i and then leaves the study for any reason (e.g. lost to follow up) t_i is counted as their censorship time.

The variance of S hat is estimated using the method of Greenwood (1926):

$$\hat{\mathrm{Var}}[\hat{S}(t)] = \hat{S}(t) \sum_{t_i \le t} \frac{d_i}{n_i (n_i - d_i)}$$

The confidence interval for the survivor function is not calculated directly using Greenwood's variance estimate because this would give impossible results (< 0 or > 1) at extremes of S. The confidence interval for S hat uses an asymptotic maximum likelihood solution by log transformation as recommended by Kalbfleisch and Prentice (1980).

The median survival time is calculated as the smallest survival time for which the survivor function is less than or equal to 0.5. A confidence interval for the median survival time is constructed using the nonparametric method due to Brookmeyer and Crowley (1982). Another confidence interval for the median survival time is constructed using a large sample estimate of the density function of the survival estimate (Andersen, 1993). If there are tied survival times then the user is advised that the Brookmeyer-Crowley limits might not be appropriate.

Mean survival time is estimated as the area under the survival curve. The estimator is based upon the entire range of data. Hosmer and Lemeshow (1999) point out that some software biases the estimate of the mean downwards by considering only the observed (uncensored) event times, and they recommend that the entire range of data is used. A large sample method is used to estimate the variance of the mean survival time, and thus to construct a confidence interval (Andersen, 1993).

The cumulative hazard function (H) is the risk of event (e.g. death) at time t, it is estimated as minus the natural logarithm of the product limit estimate of the survivor function as above (Peterson, 1977). Statistical software such as Stata (Stata Corporation, 1999) calculates the simpler Nelson-Aalen estimate (Nelson, 1972; Aalen, 1978):

$$\tilde{H}(t) = \sum_{t_i \le t} \left(\frac{d_i}{n_i} \right)$$

A Nelson-Aalen hazard estimate will always be less than an equivalent Peterson estimate and there is no substantial case for using one in favour of the other.

The variance of H hat is estimated as:

$$\hat{\mathrm{V}}\mathrm{ar}[\hat{H}(t)] = \frac{\hat{\mathrm{V}}\mathrm{ar}[\hat{S}(t)]}{\hat{S}(t)^2}$$

Life table

The Berkson and Gage method is used to construct a simple life table (actuarial table) that displays the survival experience of a cohort (Berkson and Gage, 1950; Armitage and Berry, 1994; Altman, 1991; Lawless, 1982; Kalbfleisch and Prentice, 1980; Le, 1997). The table is constructed by the following definitions:

Interval For a full life table this is ages in single years.
For an abridged life table this is ages in groups.
For a Berkson and Gage survival table this is the survival times in intervals. [t]

Deaths Number of individuals who die in the interval. [d_x]

Withdrawn Number of individuals withdrawn or lost to follow up in the interval. [w_x]

At Risk Number of individuals alive at the start of the interval. [n_x]

Adj. at risk Adjusted number at risk (half of withdrawals of current interval subtracted). [n'_x]

P(death) Probability that an individual who survived the last interval will die in the current interval. [q_x]

P(survival) Probability that an individual who survived the last interval will survive the current interval. [p_x]

% Survivors (l_x) Probability of an individual surviving beyond the current interval.

Proportion of survivors after the current interval.

Life table survival rate.

Var(l_x %) Estimated variance of l_x.

***% CI for l_x %** *% confidence interval for l_x %.

$$n'_x = n_x - \frac{w_x}{2}$$

$$q_x = \frac{d_x}{n'_x}$$

$$p_x = 1 - q_x$$

$$l_x = \prod_{i=0}^{x-1} p_i$$

- where l_x is the product of all p_x before x.

The confidence interval for l_x is not a simple application of the estimated variance, instead it uses a maximum likelihood solution from an asymptotic distribution by the transformation of l_x suggested by Kalbfleisch and Prentice (1980). This treatment of l_x avoids impossible values (i.e. >1 or <0).

Log-rank and Wilcoxon

Log-rank and Wilcoxon type tests (null hypothesis - risk of death/event same in all groups) are provided for comparing two or more survival curves where some of the observations may be censored and where the overall grouping may be stratified (Tarone and Ware, 1977; Kalbfleisch and Prentice, 1980; Cox and Oakes, 1984; Hosmer and Lemeshow, 1999).

The general test statistic is calculated around a hypergeometric distribution of the number of events at distinct event times:

$$e_{ij} = \frac{n_{ij} d_j}{n_j}$$

$$\text{diagonal}(\hat{\mathbf{V}}_j)_{ii} = \frac{n_{ij}(n_j - n_{ij}) d_j (n_j - d_j)}{n_j^2 (n_j - 1)}, i = 1,2...,k-1$$

$$\text{off diagonal}(\hat{\mathbf{V}}_j)_{ii} = \frac{n_{ij} n_{lj} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}, i,l = 1,2...,k-1, k \neq l$$

$$\mathbf{U}_j = \sum_{i=1}^{k} w_j (d_{ij} - e_{ij})$$

$$\hat{\mathbf{V}}_\mathbf{w} = \mathbf{w}\hat{\mathbf{V}}\mathbf{w}$$

$$\chi^2_{k-1} = \mathbf{U}' \mathbf{V}_\mathbf{w}^{-1} \mathbf{U}$$

$$\chi^2_{k-1}\, trend = \frac{(\mathbf{c}'\mathbf{U})^2}{\mathbf{c}'\hat{\mathbf{V}}_\mathbf{w}^{-1}\mathbf{c}}$$

- where the weight w_j for the log-rank test is equal to 1, and w_j for the generalized Wilcoxon test is n_i (Gehan-Breslow method); for the Tarone-Ware method w_j is the square root of n_i; and for the Peto-Prentice method w_j is the Kaplan-Meier survivor function multiplied by (n_i divided by n_i *+1*). e_{ij} is the expectation of death in group *i* at the j^{th} distinct observed time where d_j events/deaths occurred. n_{ij} is the number at risk in group *i* just before the j^{th} distinct observed time. The test statistic for equality of survival across the *k* groups (populations sampled) is approximately chi-square distributed on *k-1* degrees of freedom. The test statistic for monotone trend is approximately chi-square distributed on 1 degree of freedom. **c** is a vector of scores that are either defined by the user or allocated as 1 to *k*.

Variance is estimated by the method that Peto et al. (1977) refer to as "exact".

In the absence of censorship, the methods presented here reduce to a Mann-Whitney (two sample Wilcoxon) test for two groups of survival times or to a Kruskal-Wallis test for more than two groups of survival times.

The user is informed that Peto's log-rank test is generally the most appropriate method but that the Prentice modified Wilcoxon test is more sensitive when the ratio of hazards is higher at early survival times than at late ones (Peto and Peto,

1972; Kalbfleisch and Prentice, 1980). The log-rank test is similar to the Mantel-Haenszel test and some authors refer to it as the Cox-Mantel test (Mantel and Haenszel, 1959; Cox, 1972).

The user can opt to stratify the groups specified and to test the significance of the stratification (Armitage and Berry, 1994; Lawless, 1982; Kalbfleisch and Prentice, 1980). The stratified test statistic is expressed as:

$$\chi^2_{k-1} = \left(\sum \mathbf{U}\right)' \left(\sum \hat{\mathbf{V}}_{\mathbf{w}}\right)^{-1} \left(\sum \mathbf{U}\right)$$

- where the statistics defined above are calculated within strata and then summed across strata prior to the generalised inverse and transpose matrix operations.

A choice of three different weighting methods is given for the generalised Wilcoxon test, these are Peto-Prentice, Gehan and Tarone-Ware. The Peto-Prentice method is generally more robust than the others are but the Gehan statistic is calculated routinely by many statistical software packages (Breslow, 1974; Tarone and Ware, 1977; Kalbfleisch and Prentice, 1980; Miller, 1981; Hosmer and Lemeshow 1999).

An approximate confidence interval for the log hazard-ratio is calculated using the following estimate of standard error (*se*):

$$se = \sqrt{\sum_{i=1}^{k} \frac{1}{e_{ij}}}$$

- where e_i is the extent of exposure to risk of death (sometimes called expected deaths) for group i of k at the j^{th} distinct observed time (Armitage and Berry, 1994). The user is informed that Cox regression gives a more accurate estimate of this statistic.

With more than two groups, a variant of the log-rank test for trend is calculated. If the user does not specify scores then they are allocated as *1,2,3 ... n* in group order (Armitage and Berry, 1994; Lawless, 1982; Kalbfleisch and Prentice, 1980; Hosmer and Lemeshow 1999).

Wei-Lachin

The log-rank and generalised Wilcoxon methods are extended here for the comparison of multivariate survival data. Wei and Lachin's multivariate tests are calculated for the case to two multivariate distributions, and the intermediate univariate statistics are given. The algorithm used for the method is that given by Makuch and Escobar (1991).

The general univariate statistic for comparing the time to event (of component type k out of m multivariate components) of the two groups is calculated as:

$$T_k = \frac{\sum_{j=1}^{n_1} w_j \Delta_{1kj} e_{2k} - \sum_{j=1}^{n_2} w_j \Delta_{2kj} e_{1k}}{\sqrt{n}}$$

$$e_{ik} = \frac{r_{ik}}{r_{1k} + r_{2k}}$$

- where n_1 is the number of event times per component in group 1; n_2 is the number of event times per component in group 2; n is the total number of event times per component; r_{ik} is the number at risk at time t(i) in the k^{th} component; Δ is equal to 0 if an observation is censored or 1 otherwise; e_{ik} is the expected proportion of events in group i for the k^{th} component; and w_j is equal to 1 for the log-rank method or $(r_{1k}+r_{2k})/n$ for the Gehan-Breslow generalised Wilcoxon method.

The univariate statistic for the k^{th} component of the multivariate survival data is calculated as:

$$WL_k = \frac{T_k}{\sqrt{\hat{\sigma}_{kk}}}$$

- where σ_{kk} caret is the k^{th} diagonal element of the estimated variance-covariance matrix that is calculated as described by Makuch and Escobar (1991).

An omnibus test that the two multivariate distributions are equal is calculated as:

$$WL_0 = \mathbf{T'}\boldsymbol{\Sigma}^{-1}\mathbf{T}$$

- where **T'** is the transpose of the vector of univariate test statistics and $\mathbf{S^{-1}}$ is the generalised inverse of the estimated variance-covariance matrix.

A stochastic ordering test statistic is calculated as:

$$WL_1 = \frac{\sum_{k=1}^{m} T_k}{\sqrt{\sum_{k=1}^{m}\sum_{l=1}^{m} \hat{\sigma}_{kl}}}$$

The *WL* test statistics above are asymptotically normally distributed.

Cox regression

The Cox proportional hazards model is based around a multiplicative effect on the hazard function, of factors affecting survival:

$$h(t \mid \mathbf{x}) = h_0(t)e^{\mathbf{x'\beta}}$$

- where **x** is a vector of regressor variables, **β** is a vector of unknown parameters, and $h_0(t)$ is the baseline hazard function for a subject with **x** = 0.

Maximum likelihood estimates for **β** are found by Newton-Raphson iteration with log partial/marginal likelihood function as:

$$l = \ln L(\boldsymbol{\beta}) = \sum_{j=1}^{m}\sum_{i=1}^{k_j} \mathbf{S_{ji}'}\boldsymbol{\beta} - \sum_{j=1}^{m}\sum_{i=1}^{k_j} d_{ji} \ln\left[\sum_{l \in R_{ji}} e^{\mathbf{x_l'\beta}}\right]$$

- where l is the log partial likelihood, R_{ji} is the risk set (subjects alive and uncensored just prior to t_{ji}) for the i^{th} event time in the j^{th} stratum and S_{ji} is the sum of the covariates for subjects dying (event) at the i^{th} event time in the j^{th} stratum.

Derivatives for the optimisation and variance estimation are as defined by Lawless (1982) and Kalbfleisch and Prentice (1980). The user had some control over convergence criteria in that they can specify a target precision for the estimates

and a ratio for the proportionality constant (exponent of the sum of covariates multiplied by their estimated parameters at an observed time) beyond which relevant data are split into separate strata automatically (Bryson and Johnson, 1981). Ties are handled by Breslow's approximation (Breslow, 1974). A matching quasi-Newton method is used to approximate the variances of the estimated parameters (Lawless, 1982).

Cox-Oakes residuals and influence statistics are calculated as described by Cox and Oakes (1984). Cox-Snell, Martingale and deviance residuals are calculated as described by Collett (1994).

Survival and cumulative hazard rates are calculated for the mean covariate value and the baseline at each event time:

$$\hat{S}_0(t) = \prod_{i:(t_i<t)} \hat{\alpha}_i$$

$$\sum_{l \in D_i} \frac{\hat{\theta}_l}{1-\hat{\alpha}_i^{\hat{\theta}_l}} = \sum_{l \in R_i} \hat{\theta}_l$$

$$\hat{\theta}_l = e^{\mathbf{x}'\boldsymbol{\beta}}$$

- where the caret denotes estimated values and $S_0(t)$ is the survivor function. The iterative maximum likelihood method above is used when there are tied event times and the closed form below is used when there is only one death/event at a particular survival time (Kalbfleisch and Prentice, 1973):

$$\hat{\alpha}_i = \left[1 - \frac{\hat{\theta}_l}{\sum_{l \in R_i} \hat{\theta}_l} \right]^{\hat{\theta}_i^{-1}}$$

The cumulative hazard function is estimated as -log of the estimated survivor function.

Meta-analysis

Different strata of observations are investigated both individually and pooled by methods of meta-analysis in this section. Both fixed and random effects models are considered and the user is informed that there are no universally agreed rules on which model is most appropriate in different circumstances (Fleiss and Gross, 1991; Sahai and Kurshid, 1996; DerSimonian and Laird 1985).

For each design, a Q ("combinability") statistic is given with its associated probability on k (number of strata) minus one degrees of freedom. The user is warned that this has low power as a strict test of homogeneity; it is included as a part of the DerSimonian-Laird random effects analysis (DerSimonian and Laird 1985).

Odds ratio

The odds ratio is examined for each stratum of and overall for a group of related studies.

For a single stratum the odds ratio is estimated as follows:

		Exposed	Non-Exposed
OUTCOME:	Cases	a	b
	Non-cases	c	d

Point estimate of the odds ratio = (ad)/(bc)

For each table, the observed odds ratio is calculated with an exact confidence interval (Thomas, 1971; Sahai and Kurshid, 1996). With very large numbers (n > 100000), the Cornfield approximation is used to calculate the confidence interval otherwise the exact (Gart) method is used (Sahai and Kurshid, 1996; Fleiss, 1979).

The Mantel-Haenszel method is used to estimate the pooled odds ratio for all strata under the assumption of a fixed effects model:

$$\hat{O}R_{MH} = \frac{\sum_{i=1}^{k}\left(\frac{a_i d_i}{n_i}\right)}{\sum_{i=1}^{k}\left(\frac{b_i c_i}{n_i}\right)}$$

- where $n_i = a_i+b_i+c_i+d_i$.

A confidence interval for the Mantel-Haenszel odds ratio is calculated using the Robins, Breslow and Greenland variance formula (Robins et al., 1986) or by the method of Sato (1990) if the estimate of the odds ratio can not be determined. A chi-square test statistic is given with its associated probability that the pooled odds ratio is equal to one.

Peto odds ratio

Peto's odds ratio is examined for each stratum of and overall for a group of related studies.

For a single stratum Peto's odds ratio is estimated as follows (Yusuf et al. 1985):

		Exposed	Non-Exposed
OUTCOME:	Cases	a	b
	Non-cases	c	d

$$\hat{\psi} = \exp\left(\frac{O-E}{V}\right)$$

$$O = a$$

$$E = (a+b)(a+c)/n$$

$$V = \frac{(a+b)(c+d)(a+c)(b+d)}{n^2(n-1)}$$

$$z = \frac{O-E}{\sqrt{V}}$$

$$CI = \exp\left(\frac{(O-E) \pm z_{\alpha/2}\sqrt{V}}{V}\right)$$

- where psi hat is the Peto odds ratio, n = a+b+c+d, CI is the 100(1-α)% confidence interval and z is a quantile from the standard normal distribution. V is both weighting factor and variance for the difference between observed and expected a, O-E.

The pooled Peto odds ratio and its confidence interval are calculated as follows:

$$\hat{\psi}_{pool} = \exp\left(\frac{\sum_{i=1}^{k}(O_i - E_i)}{\sum_{i=1}^{k} V_i}\right)$$

$$CI_{pool} = \exp\left(\frac{\sum_{i=1}^{k}(O_i - E_i) \pm z_{\alpha/2}\sqrt{\sum_{i=1}^{k} V_i}}{\sum_{i=1}^{k} V_i}\right)$$

The user is informed that the Peto odds ratio is an alternative to the usual Mantel-Haenszel method for pooling odds ratios across the strata of fourfold tables, and that it is not mathematically equal to the classical odds ratio. The user is warned that the Peto odds ratio can cause avoidable biases in analysis and that it should therefore only be used under the expert guidance of a statistician (Greenland and Salvan, 1990; Fleiss, 1993).

The Q ("combinability") statistic is calculated as:

$$Q = \sum_{i=1}^{k} \frac{\left(\frac{O_i - E_i}{V_i}\right)^2}{V_i}$$

The user is warned that the combinability test has low power as a strict test of homogeneity (Fleiss and Gross, 1991; Sahai and Kurshid, 1996).

Relative risk

Relative risk is examined by stratum and overall for a group of related studies.

For a single stratum relative risk is defined as follows:

		Exposed	Non-Exposed
OUTCOME:	Cases	a	b
	Non-cases	c	d

Relative risk = [a/(a+c)] / [b/(b+d)]

For each table the observed relative risk is displayed with a near exact confidence interval. The iterative methods for ratios of binomial probabilities described by Gart and Nam are used in this procedure (Gart and Nam 1988; Sahai and Kurshid, 1996).

The Mantel-Haenszel type method of Rothman and Boice (Rothman and Greenland, 1998) is used to estimate the pooled risk ratio for all strata under the assumption of a fixed effects model:

$$R\hat{R}_{MH} = \frac{\sum_{i=1}^{k}\left(\frac{b_i + d_i}{n_i}\right)a_i}{\sum_{i=1}^{k}\left(\frac{a_i + c_i}{n_i}\right)b_i}$$

- where $n_i = a_i+b_i+c_i+d_i$.

A confidence interval for the pooled relative risk is calculated using the Greenland-Robins variance formula (Greenland and Robins, 1985). A chi-square test statistic is given with associated probability of the pooled relative risk being equal to one.

Risk difference

Relative difference is examined by stratum and overall for a set of related studies.

For a single stratum risk difference is defined as follows:

		Exposed	Non-Exposed
OUTCOME:	Cases	a	b
	Non-cases	c	d

Risk difference = [a/(a+c)] - [b/(b+d)]

For each table, the observed risk difference is displayed with a near exact confidence interval. The iterative method of Miettinen and Nurminen is used to construct the confidence interval for the difference between the unpaired proportions that constitute the risk difference (Mee, 1984; Anbar, 1983; Gart and Nam, 1990; Miettinen and Nurminen, 1985; Sahai and Kurshid, 1991).

The Mantel-Haenszel type method of Greenland and Robins (Greenland and Robins, 1985; Sahai and Kurshid, 1991) is used to estimate the pooled risk difference for all strata under the assumption of a fixed effects model:

$$\hat{RD}_{MH} = \frac{\sum_{i=1}^{k}\left[a_i\left(\frac{b_i + d_i}{n_i}\right) - b_i\left(\frac{a_i + c_i}{n_i}\right)\right]}{\sum_{i=1}^{k}\left(\frac{(a_i + c_i)(b_i + d_i)}{n_i}\right)}$$

- where $n_i = a_i+b_i+c_i+d_i$.

A confidence interval for the pooled risk difference is calculated using the Greenland-Robins variance formula (Greenland and Robins, 1985). A chi-square test statistic is given with associated probability of the pooled risk difference being equal to zero.

Effect size

There are a number of different statistical methods for estimating effect size; the two methods used in this procedure are *g* (modified Glass statistic with pooled sample standard deviation) and the unbiased estimator *d* (Hedges and Olkin, 1985). *g* and *d* are calculated as follows:

$$d = g[J(N-2)]$$

$$g = \frac{\mu_e - \mu_c}{\sigma_{pooled}}$$

$$\sigma_{pooled} = \sqrt{\frac{\sigma_e^2(n_e - 1) + \sigma_c^2(n_c - 1)}{N - 2}}$$

$$J(m) = \frac{\Gamma(m/2)}{\Gamma[(m-1)/2]\sqrt{m/2}}$$

- where n_e is the number in the experimental group, n_c is the number in the control group, μ_e is the sample mean of the experimental group, μ_c is the sample mean of the control group, σ_e is the sample standard deviation for the experimental group, μ_c is the sample standard deviation for the control group, $N = n_e + n_c$, J_m is the correction factor given m and Γ is the gamma function.

For each study *g* is given with an exact confidence interval and *d* is given with an approximate confidence interval. An iterative method based on the non-central t distribution is used to construct the confidence interval for *g* (Hedges and Olkin, 1985).

The pooled mean effect size estimate (*d*+) is calculated using direct weights defined as the inverse of the variance of *d* for each stratum. An approximate confidence interval for *d*+ is given with a chi-square statistic and probability of this pooled effect size being equal to zero (Hedges and Olkin, 1985).

The user is given the option to base effect size calculations on weighted mean difference (a non-standardized estimate unlike *g* and *d*) as described in the Cochrane Collaboration Handbook (Mulrow and Oxman, 1996).

Incidence rate

Incidence rates are examined by stratum and overall for a group of related studies.

Person-time is the sum of times that subjects in a sample have been studied for:

	Exposed		Not exposed	
	Cases	Person-time	Cases	Person-time
stratum 1	a_1	pt_{e1}	b_1	pt_{n1}
	.	.	.	.
stratum k	a_k	pt_{ek}	b_k	pt_{nk}

For each stratum:

Incidence rate difference = IRD = $[a/pt_e] - [b/pt_n]$

Incidence rate ratio = IRR = $[a/pt_e] / [b/pt_n]$

The user is given the option of IRD or IRR based meta-analysis. For each stratum, either IRD (with approximate confidence interval) or IRR (with exact confidence interval) is calculated. Pooled estimates for IRD or IRR are given for both fixed and random effects models (Sahai and Kurshid, 1996; Ioannidis et al., 1995; Rothman and Monson 1983; DerSimonian and Laird, 1986).

Pooled incidence rate difference for fixed effects is estimated as follows:

$$\hat{IRD} = \frac{\sum_{i=1}^{k}\left[W_i\left(\frac{a_i}{pt_{ei}} - \frac{b_i}{pt_{ni}}\right)\right]}{\sum_{i=1}^{k} W_i}$$

$$W_i = \frac{pt_{ei}^2 pt_{ni}^2}{a_i pt_{ni}^2 - b_i pt_{ei}^2}$$

- where the weight, W_i, is the inverse of the estimated variance.

Pooled incidence rate ratio for fixed effects is estimated as follows:

$$\hat{IRR} = \exp\left\{ \frac{\sum_{i=1}^{k}\left[W_i \log\left(\frac{a_i}{pt_{ei}} / \frac{b_i}{pt_{ni}} \right)\right]}{\sum_{i=1}^{k} W_i} \right\}$$

$$W_i = \frac{a_i b_i}{a_i + b_i}$$

- where the weight, W_i, is the inverse of the estimated variance for the log transformed statistic.

Graphics

A vector coordinate based system, as opposed to a bitmap pixel based system, is used for constructing charts. Microsoft Windows metafile format is the vector system used. This enables the user to scale charts without loss of proportion.

A simple two-dimensional style is adopted in order to minimise loss of detail when charts are scaled to a small size for publication.

A general, object-oriented charting object is available to the user. Most of the graphical output, however, is focused upon chart formats that are difficult to find in commonly used spreadsheet and charting software. Examples of such charts are ladder plots, box and whisker plots and population pyramids.

The Receiver Operating Characteristic (ROC) plot function provides statistical results in addition to a graphical plot of the ROC curve. The curve is constructed by enumeration of counts derived from observations classified into two groups as observed and as calculated when the cut-off for membership of one group is varied (Altman 1991). The user is presented with a "cut-off calculator" that shows the change in sensitivity, specificity and predictive values as the cut-off point is varied. The area under the ROC curve is calculated directly by extended trapezoidal rule (Press et al. 1992) and indirectly by a Wilcoxon type method with a confidence interval (Hanley and McNeil 1982).

Sustainable development and distribution

Development of the software is sustained by sales of licences to use the software.

From 1990 to 1996, the first versions (Arcus, Arcus Professional and then Arcus Pro-Stat) of the software for the DOS platform were distributed as shareware. Shareware is software that is distributed freely to users who are asked to pay a registration fee if they continue to use the software beyond a specified trial period. A printed instruction book plus the latest version of the software was sent to all users who registered. The number of unregistered users was not measurable.

Revenue from Arcus shareware funded computer hardware and software development tools for the next phase of research and development. The product of this work was Arcus QuickStat (Biomedical) software for the 16 bit Microsoft Windows 3.11 operating system. Arcus QuickStat was published in 1996 by Addison Wesley Longman Ltd.. Licensed users of this software received printed documentation. In 1997, a web site was established to support users of Arcus QuickStat. Updates to the software were distributed via the web site. Statistical calculations were added and/or improved through incremental updates.

Royalties from Arcus QuickStat were used to fund new computer hardware, software development tools and Internet facilities to help build the next generation of Arcus software for 32 bit Windows platforms. This was launched as StatsDirect in 1999.

All of the algorithms in StatsDirect were added afresh or re-written to compute in 64-bit precision, previous Arcus software used 32-bit precision. The documentation for StatsDirect was written in the form of a basic statistical knowledge base and provided in electronic form only.

A web site was established at www.statsdirect.com and www.statsdirect.co.uk for the distribution of software and the support of users. No publisher or other third party is involved in the distribution of StatsDirect; there is a direct link from user

to author via the web site. An ethical pricing policy is operated; this provides low cost licences for students, academics and people from the developing world.

Evaluative feedback was facilitated by sending the software to reviewers and statistical authors, and by including links from later versions of the software to Internet based discussion areas.

Results

"On two occasions I have been asked [by members of Parliament], 'Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?' I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question."
Babbage, Charles (1792-1871)

Numerical validation

For each set of validation results below, either the test data are quoted or reference is made to them in the test workbook that is provided in the electronic part of this thesis.

Standard normal distribution

$z_{0.001}$ = -3.09023230616781
Lower tail P(z= -3.09023230616781) = 0.001
$z_{0.25}$ = -0.674489750196082
Lower tail P(z= 0.674489750196082) = 0.25
z_{1E-20} = -9.26234008979841
Lower tail P(z= -9.26234008979841) = 9.99999999999962E-21

Complete agreement is found with published tables (Pearson and Hartley, 1970). The first two results above agree to 15 decimal places with the reference data of Wichura (1988). The extreme value (lower tail P of 1E-20) evaluates correctly to 14 decimal places.

Student's t distribution

Syntax is $t_{P,\ \text{degrees of freedom}}$:
Two tailed $t_{0.001,\ 1}$ = 636.61924876872 (upper tail P = 0.0005)
Two tailed $t_{0.05,\ 2}$ = 4.30265272974946 (upper tail P = 0.025)
Two tailed $t_{0.01,\ 60}$ = 2.66028302884958 (upper tail P = 0.005000000000072)
Two tailed $t_{0.05,\ 120}$ = 1.97993040508242 (upper tail P = 0.025000000000002)

Complete agreement is found with published tables (Pearson and Hartley, 1970). Maximum truncation error observed was beyond the 13th decimal place.

F (variance ratio) distribution

Syntax is $F_{P,\text{numerator degrees of freedom, denominator degrees of freedom}}$:

$F_{0.001, 8, 1} = 598144.156249954$ (upper tail P = 0.001)

$F_{0.05, 2, 2} = 19$ (upper tail P = 0.05)

$F_{0.01, 1, 12} = 9.33021210316856$ (upper tail P = 0.01)

$F_{0.005, 8,1} = 238.882694802524$ (upper tail P = 0.005)

Complete agreement is found with published tables (Pearson and Hartley, 1970) and with critical test values (Berry et al., 1990). Calculated areas and quantiles agreed to the limit of 64 bit decimal precision.

Chi-square distribution

Syntax is $\chi^2_{P, \text{ degrees of freedom}}$:

$\chi^2_{0.05, 20} = 31.4104328442309$ (upper tail P = 0.05)

$\chi^2_{0.01, 2} = 9.21034037197618$ (upper tail P = 0.01)

$\chi^2_{0.001, 100} = 149.449252779038$ (upper tail P = 0.001)

$\chi^2_{0.000916828861456, 60} = 99.9999999999999$ and 100 (upper tail P = 0.000916828861456)

Complete agreement is found with published tables (Pearson and Hartley, 1970). Maximum truncation error observed was beyond the 13th decimal place.

Studentized range distribution

Syntax is $Q_{P, \text{ degrees of freedom, number of samples}}$:

$Q_{0.05, 12, 3} = 3.772929$ (upper tail P = 0.05)

$Q_{0.05, 5, 4} = 5.2183249$ (upper tail P = 0.05)

$Q_{0.01, 30, 10} = 5.7562549$ (upper tail P = 0.01)

$Q_{0.01, 8, 9} = 7.6803324$ (upper tail P = 0.01)

Complete agreement is found with published tables (Pearson and Hartley, 1970). Maximum truncation error observed was beyond the 7th decimal place.

Binomial distribution

Syntax is upper tail $P(\leq r)_{\text{binomial parameter p, n Bernoulli trials, r successes}}$:

$P_{0.5, 8, 2} = 0.14453125$

$P_{0.5, 11, 5} = 0.500000000000001$

$P_{0.5, 14, 0} = 0.00006103515625$

$P_{0.7, 17, 12} = 0.611310356950235$

Complete agreement is found with published tables (Conover, 1999; Pearson and Hartley, 1970). Maximum truncation error observed was beyond the 14th decimal place.

Poisson distribution

Syntax is upper tail $P(\leq k)_{\text{k events, mean } \mu}$:

$P_{5, 10.5} = 0.050380451088936$

$P_{3, 0.4} = 0.999223748623793$

$P_{0, 7.5} = 0.000553084370148$

$P_{25, 16} = 0.986881437112416$

Complete agreement is found with published tables (Hogg and Tanis, 1993; Pearson and Hartley, 1970). Maximum truncation error observed was beyond the 14th decimal place.

Kendall's test statistic and tau distribution

Syntax is upper tail $P_{\text{test statistic, sample size}}$:

$P_{24, 9} = 0.006332671957672$

$P_{129, 30} = 0.010726438996435$

$P_{19, 10} = 0.054156746031746$

$P_{4, 4} = 0.166666666666667$

Complete agreement is found with published tables (Conover 1999; Hollander and Wolfe, 1999). Maximum truncation error observed was beyond the 14th decimal place for sample sizes of less than or equal to 50 and beyond the 7th decimal place for sample sizes of greater than 50. Data are assumed to have been ranked without ties.

Hotelling's test statistic and Spearman's rho distribution

Syntax is upper tail $P_{\text{test statistic, sample size:}}$

$P_{2, 5} = 0.008333333333333$

$P_{8, 6} = 0.029166666666667$

$P_{66, 11} = 0.00936419340051$

$P_{934, 20} = 0.099703213545821$

Complete agreement is found with published tables (Conover 1999; Hollander and Wolfe, 1999). Maximum truncation error observed was beyond the 14th decimal place for sample sizes of less than or equal to 9 and the 4th decimal place for sample sizes of greater than 9.

Non-central t distribution

Syntax is $T_{\text{P, degrees of freedom, non-centrality parameter:}}$

$T_{0.05, 4, 4} = 10.1155420333576$ (upper tail P = 0.050000000000001)

$T_{0.01, 9, 3} = 7.70491977062513$ (upper tail P = 0.01)

$T_{0.1, 30, 3} = 4.517393111753$ (upper tail P = 0.1)

$T_{0.01, 2, 2} = 22.251180427131$ (upper tail P = 0.01)

Complete agreement is found with published tables (Owen, 1965). Maximum truncation error observed was beyond the 14th decimal place.

Sign test

Data are from Altman (1991): 11 binomial observations with 9 in one direction.

Sign test

For 11 observations with 9 on one side:

Cumulative probability (two sided) = 0.06543
(one sided) = 0.032715

Normal approximate Z = 1.809068
Two sided P = 0.0704
One sided P = 0.0352

Exact (Clopper-Pearson) 95% confidence interval for the proportion:
Lower Limit = 0.482244
Proportion = 0.818182
Upper Limit = 0.977169

Fisher's exact test

Data are from Armitage and Berry (1994).

Fisher's exact test

Input table:
4 16
1 21

Arranged table and totals:
4 1 5
16 21 37
20 22 42
Expectation of A = 2.380952

One sided (upper tail) P = .1435 (doubled one sided P = 0.2871)
Two sided (by summation) P = 0.1745

Expanded Fisher's exact test

Data are from Armitage and Berry (1994).

Fisher's exact test (expanded)

Input table:

4	16
1	21

Arranged table and totals:

4	1	5
16	21	37
20	22	42

Expectation of A = 2.38095238095238

A	Lower Tail	Individual P	Upper Tail
0	1.000000000000000	0.030956848030019	0.030956848030019
1	0.202939337085679	0.171982489055660	0.969043151969981
2	0.546904315196998	0.343964978111320	0.797060662914321
3	0.856472795497186	0.309568480300188	0.453095684803002
4	0.981774323237738	0.125301527740552	0.143527204502814
5	1.000000000000000	0.018225676762262	0.018225676762262

One sided (upper tail) P = 0.143527 (doubled: 0.287054)

Two sided (by summation) P = 0.174484

McNemar and exact (Liddell) test

Data are from Armitage and Berry (1994).

McNemar and exact (Liddell) test for a paired fourfold table

Input table:

20	12
2	16

Uncorrected Chi² = 7.142857 (1 DF) P = 0.0075
Yates' continuity corrected Chi² = 5.785714 (1 DF) P = 0.0162

After Liddell (1983):
Point estimate of relative risk (R') = 6
Exact 95% confidence interval = 1.335744 to 55.197091

F = 4
Two sided P = 0.0129

R' is significantly different from unity

Exact confidence limits for 2 by 2 odds

Data are from Thomas (1971).

Exact confidence limits for 2 by 2 odds (Gart)

Input table:

10	3
2	15

Re-arranged Table and totals:

15	2	17
3	10	13
18	12	30

Confidence limits with 2.5% lower tail area and 2.5% upper tail area:

Observed odds ratio = 25
Confidence interval = 2.753383 to 301.462141

Reciprocal = 0.04
Confidence interval = 0.003317 to 0.36319

Chi-square test (2 by 2)

Data are from Armitage and Berry (1994) treated as case-control study.

Chi-square test (2 by 2)

Observed values and totals:

41	216	257
64	180	244
105	396	501

Expected values:

53.862275	203.137725
51.137725	192.862275

Uncorrected $Chi^2 = 7.978869$ P = 0.0047

Yates-corrected $Chi^2 = 7.370595$ P = 0.0066

Coefficient of contingency: V = -0.126198

Odds ratio analysis

Using the Woolf (logit) method:

Odds Ratio = 0.533854

95% CI (logit method) = 0.344118 to 0.828206

Using Gart's method for a 95% confidence interval:

Odds ratio = 0.533854 (reciprocal 1.873171)

Lower limit = 0.334788 (reciprocal limit 1.181625)

Upper limit = 0.846292 (reciprocal limit 2.986966)

Chi-square test (2 by k)

Data are from Armitage and Berry (1994).

Chi-square test (2 by k)

	Successes	Failures	Total	Per cent
Observed	19	497	516	3.68
Expected	26.575107	489.424893		
Observed	29	560	589	4.92
Expected	30.334764	558.665236		
Observed	24	269	293	8.19
Expected	15.090129	277.909871		
Observed	72	1326	1398	5.15
Expected	72	1326		
Total	144	2652	2796	5.15

Total Chi^2 = 7.884843 |Chi| = 2.807996 (3 DF) P = 0.0485

Chi^2 for linear trend = 1.379316 |Chi| = 1.174443 (1 DF) P = 0.2402

Remaining Chi^2 (non-linearity) = 6.505527 (2 DF) P = 0.0387

Chi-square test (r by c)

Data are from Armitage and Berry (1994).

Chi-square test (r by c)

Observed	17	9	8	34	1
Expected	13.909091	10.818182	9.272727		
DChi²	0.686869	0.305577	0.174688		
Observed	6	5	1	12	2
Expected	4.909091	3.818182	3.272727		
DChi²	0.242424	0.365801	1.578283		
Observed	3	5	4	12	3
Expected	4.909091	3.818182	3.272727		
DChi²	0.742424	0.365801	0.161616		
Observed	1	2	5	8	4
Expected	3.272727	2.545455	2.181818		
DChi²	1.578283	0.116883	3.640152		
Total	27	21	18	66	
Score	1	2	3		

TOTAL number of cells = 12

WARNING: 9 out of 12 cells have EXPECTATION < 5

INDEPENDENCE

Chi-square = 9.9588	DF = 6	P = 0.1264
G-square = 10.186039	DF = 6	P = 0.117
Fisher's exact (two sided)		P = 0.1426

ANOVA

Chi-square for equality of mean column scores = 5.696401

DF = 2 P = 0.0579

LINEAR TREND

Sample correlation (r) = 0.295083

Chi-square for linear trend (M^2) = 5.6598

DF = 1 P = 0.0174

ASSOCIATION

Phi = 0.388447

Pearson's contingency = 0.362088

Cramér's V = 0.274673

Woolf chi-square statistics

Data are from Armitage and Berry (1994) and given in the columns "Experimental group size", "Experimental responders", "Control group size" and "Control responders" of the "test" workbook.

Woolf chi-square analysis of 2 by 2 series

For combined tables without Haldane correction:

Number of Tables = 10

Mean log odds ratio = 1.508337 giving odds ratio = 4.519207

Variance of mean log odds ratio = 0.008985 standard error = 0.094789

Approximate 95% CI for mean log odds ratio = 1.322554 to 1.694119

Giving odds ratio of 3.752994 to 5.441851

Chi^2 for expected log odds ratio = 0 is 253.21084 Chi = 15.9126 P < 0.0001

Chi^2 for heterogeneity = 6.634076 DF = 9 P = 0.6752

For combined tables with Haldane correction:

Number of tables = 10

Mean log odds ratio = 1.506344 giving odds ratio = 4.510211

Variance of mean log odds ratio = 0.00893 standard error = 0.0945

Approximate 95% CI for mean log odds ratio = 1.321127 to 1.691561

Giving odds ratio of 3.747641 to 5.427948

Chi^2 for expected log odds ratio = 0 is 254.086475 Chi = 15.94009 P < 0.0001

Chi^2 for heterogeneity = 6.532642 DF = 9 P = 0.6857

Mantel Haenszel chi-square test

See odds ratio meta-analysis below.

Single proportion

Data are from Armitage and Berry (1994).

Single proportion

Total = 100, response = 65
Proportion = 0.65

Exact (Clopper-Pearson) 95% confidence interval = 0.548151 to 0.742706

Using null hypothesis that the population proportion equals 0.5
Binomial one sided P = 0.0018
Binomial two sided P = 0.0035

Approximate (Wilson) 95% mid-P confidence interval = 0.552544 to 0.736358

Binomial one sided mid-P = 0.0013
Binomial two sided mid-P = 0.0027

Paired proportions

Data are from Armitage and Berry (1994).

Paired proportions

Total = 50, both = 20, first only = 12, second only = 2
Proportion 1 = 0.64
Proportion 2 = 0.44
Proportion difference = 0.2

Exact two sided P = 0.0129
Exact one sided P = 0.0065

Exact two sided mid P = 0.0074
Exact one sided mid P = 0.0037

Score based (Newcombe) 95% confidence interval for the proportion difference:
0.056156 to 0.329207

Two independent proportions

Data are from Armitage and Berry (1994).

Two independent proportions

Total 1 = 257, response 1 = 41
Proportion 1 = 0.159533
Total 2 = 244, response = 64
Proportion 2 = 0.262295

Proportion difference = -0.102762

Near exact (Miettinen) 95% confidence interval = -0.17432 to -0.031588

Exact two sided (mid) P = 0.0044

Standard error of proportion difference = 0.03638
Normal deviate (Z) = -2.824689
Approximate two sided P = 0.0047
Approximate one sided P = 0.0024

Sample sizes for paired or single sample Student t tests

Complete agreement is found with published tables (Pearson and Hartley, 1970).

Sample size for a paired or single sample Student t test

Alpha = 0.05
Power = 0.8
Difference between means = 2
Standard deviation = 1.6

Estimated minimum sample size = 8 pairs

Degrees of freedom = 7

Sample sizes for unpaired two sample Student t tests

Complete agreement is found with published tables (Pearson and Hartley, 1970).

Sample size for an unpaired two sample Student t test

Alpha = 0.05
Power = 0.8
Difference between means = 2.1
Standard deviation = 2.6
Controls per experimental subject 1

Estimated minimum sample size = 26 experimental subjects and 26 controls.

Degrees of freedom = 50

Sample sizes for independent case-control studies

Data are from Armitage and Berry (1994). Complete agreement is found with published tables (Schlesselman, 1982; Casagrande et al., 1978).

Sample size for independent case-control study

Probability of exposure in controls = 0.2
Probability of exposure in subjects = 0.333333
Controls per case subject = 1
Alpha = 0.05
Power = 0.8

For uncorrected chi-square test:
N = 172 case subjects and 172 controls

For corrected chi-square and Fisher's exact tests:
N = 187 case subjects and 187 controls

Sample sizes for independent cohort studies

Data are from Armitage and Berry (1994). Complete agreement is found with published data (Dupont, 1990; Meinert, 1986; Casagrande et al., 1978).

Sample size for independent cohort study

Probability of exposure in controls = 0.25
Probability of exposure in subjects = 0.35
Controls per case subject = 1
Alpha = 0.05
Power = 0.9

For uncorrected chi-square test:
N = 440 case subjects and 440 controls

For corrected chi-square and Fisher's exact tests:
N = 460 case subjects and 460 controls

Sample sizes for matched case-control studies

Complete agreement is found with published data (Dupont, 1988).

Sample size for matched case-control study

case-control correlation = 0
probability of exposure in controls = 0.3
odds ratio = 2
controls per case subject = 1
alpha = 0.05
power = 0.8

Estimated minimum sample size = 141

Sample size for matched case-control study

case-control correlation = 0.1
probability of exposure in controls = 0.3
odds ratio = 2
controls per case subject = 1
alpha = 0.05
power = 0.8

Estimated minimum sample size = 158

Sample sizes for paired cohort studies

Complete agreement is found with published data (Dupont, 1990; Breslow and Day, 1980).

Sample size for paired cohort study

Event rate in control group = 0.1
Event rate in experimental group = 0.3
Correlation for failure between experimental and control subjects = 0
Alpha = 0.05
Power = 0.8
Estimated minimum sample size = 60

Sample size for paired cohort study

Event rate in control group = 0.1
Event rate in experimental group = 0.3
Correlation for failure between experimental and control subjects = 0.1
Alpha = 0.05
Power = 0.8

Estimated minimum sample size = 54

Sample sizes for population surveys

Complete agreement is found with published data (Colton, 1974).

Sample size for a population survey

Population estimate	250000
Population rate	5
Maximum deviation	2
Confidence level	95

Estimated minimum sample size = 456

Risk analysis (prospective)

Data are from Altman (1991).

Risk analysis (prospective)

	Characteristic factor:	
Outcome:	Present	Absent
Positive	2	33
Negative	14	58

Risk ratio (relative risk in incidence study) = 0.344697
95% confidence interval = 0.094377 to 1.040811

Risk difference = -0.237637
95% confidence interval = -0.384949 to 0.01491

Population exposure % = 14.953271
Population attributable risk % = -10.863422
Approximate 95% confidence interval = -20.865606 to -0.861239

Risk analysis (retrospective)

Data are from Altman (1991).

Risk analysis (retrospective)

	Characteristic factor:	
Outcome:	Present	Absent
Positive	255	49
Negative	93	46

Using the Woolf (logit) method:
Odds Ratio = 2.574062
95% CI (logit method) = 1.613302 to 4.106976

Using Gart's method for a 95% confidence interval:
Odds ratio = 2.574062 (reciprocal 0.388491)
Lower limit = 1.566572 (reciprocal limit 0.237356)
Upper limit = 4.213082 (reciprocal limit 0.638337)

Population exposure % = 66.906475
Population attributable risk % = 51.294336
Approximate 95% confidence interval = 34.307694 to 68.280978

Diagnostic test (2 by 2 table)

Data are from Sackett et al. (1991).

Diagnostic test analysis

	Disease / feature:		
Test:	Present	Absent	Totals
Positive	431	30	461
Negative	29	116	145
Totals	460	146	606

Prevalence (pre-test likelihood of disease) = 0.759076 = 76%

Predictive value of +ve test
(post-test likelihood of disease) = 0.934924 = 93% {change = 17%}

Predictive value of -ve test
(post-test likelihood of no disease) = 0.8 = 80% {change = 4%}
(post-test disease likelihood despite -ve test) = 0.2 = 20% {change = -56%}

Sensitivity (true positive rate) = 0.936957 = 94%
Specificity (true negative rate) = 0.794521 = 79%

Likelihood Ratio (95% confidence interval):
LR (positive test) = 4.559855 (3.364957 to 6.340323)
LR (negative test) = 0.079348 (0.055211 to 0.113307)

Likelihood ratios (2 by k table)

Data are from Sackett et al. (1991).

Likelihood ratios

Result	+ Feature	- Feature	Likelihood Ratio	95% Confidence Interval
1	97	1	54.826087	(9.923105 to 311.581703)
2	118	15	4.446377	(2.772565 to 7.31597)
3	13	26	0.282609	(0.151799 to 0.524821)
4	2	88	0.012846	(0.003513 to 0.046227)

Number needed to treat

Data are from Haynes and Sackett (1993).

Number needed to treat

Proportion of controls suffering an event = 123/607 = 0.202636
Proportion of treated suffering an event = 94/607 = 0.15486

With near exact 95% confidence intervals:

Relative risk = 0.764228 (0.598898 to 0.974221)

Relative risk reduction = 0.235772 (0.025779 to 0.401102)

Absolute risk reduction = 0.047776 (0.004675 to 0.090991)

Number needed to treat = 20.931034 (10.990105 to 213.910426)

Number needed to treat (rounded up) = 21 (11 to 214)

Kappa inter-rater agreement with two raters

Data are from Altman (1991) and given in the columns "Negative", "Weak", "Moderate", "High" and "Very high" of the "test" workbook.

General agreement over all categories (2 raters)

Cohen's kappa (unweighted)
Observed agreement = 47.38%
Expected agreement = 22.78%
Kappa = 0.318628 (se = 0.026776)
95% confidence interval = 0.266147 to 0.371109
z (for k = 0) = 11.899574
Two sided P < 0.0001
One sided P < 0.0001

Cohen's kappa (weighted by 1-Abs(i-j)/(1 - k))
Observed agreement = 80.51%
Expected agreement = 55.81%
Kappa = 0.558953 (se = 0.038019)
95% confidence interval for kappa = 0.484438 to 0.633469
z (for kw = 0) = 14.701958
Two sided P < 0.0001
One sided P < 0.0001

Scott's pi
Observed agreement = 47.38%
Expected agreement = 24.07%
Pi = 0.30701

Disagreement over any category and asymmetry of disagreement (2 raters)

Marginal homogeneity (Maxwell) chi-square = 73.013451 df = 4 P < 0.0001

Symmetry (generalised McNemar) chi-square = 79.076091 df = 10 P < 0.0001

Screening test errors

Data are from Fleiss (1981).

	DISEASE:	
	Present	Absent
TEST: +	950 (true +ve)	10 (false +ve)
-	50 (false -ve)	990 (true -ve)

Case rate is 1/1000

False result probabilities

For an overall case rate of 10 per ten thousand population tested:

Test SENSITIVITY = 95%
Probability of a FALSE POSITIVE result = 0.913163

Test SPECIFICITY = 99%
Probability of a FALSE NEGATIVE result = 0.000051

Standardized mortality ratio

Data are from Bland (1996).

Standardized Mortality Ratio

Group-specific mortality	Observed population	Expected deaths
0.000006	1080	0.006328
0.000013	12860	0.167823
0.000047	11510	0.540245
0.000162	10330	1.668326
0.000271	7790	2.113879
		Total = 4.4966004

Standardized Mortality Ratio = 3.113463
SMR (*100 as integer) = 311
Exact 95% confidence interval = 1.702159 to 5.223862 (170 to 522)

Probability of observing 14 or more deaths by chance P = 0.0002
Probability of observing 14 or fewer deaths by chance P > 0.9999

Incidence rate analysis

Data are from Stampfer et al. (1985).

Incidence rate analysis

	Exposure:		
Outcome:	Exposed	Non-exposed	Total
Cases	30	60	90
Person-time	54308.7	51477.5	105786.2

Exposed incidence rate = 0.000552
Non-exposed incidence rate = 0.001166

Incidence rate difference = -0.000613
approximate 95% confidence interval = -0.000965 to -0.000261

chi-square = 11.678635 P = 0.0006

Incidence rate ratio = 0.473934
exact 95% confidence interval = 0.295128 to 0.746416

Basic descriptive statistics

The data are 100 measurements of the speed (millions of meters per second) of light in air recorded by Michelson in 1879 (Dorsey, 1944). The American National Institute of Standards and Technology use these data as part of the Statistical Reference Datasets for testing statistical software (McCullough and Wilson, 1999; www.nist.gov.itl/div898/strd). The data are given in the column "Michelson" of the "test" workbook.

Extended display of precision (12 decimal places) was used:

Descriptive statistics

Variables	Michelson
Valid Data	100
Missing Data	0
Mean	299.8524
Variance	0.006242666667
SD	0.079010547819
SEM	0.007901054782
Lower 95% CL	299.836722593166
Upper 95% CL	299.868077406834
Geometric Mean	299.852389694496
Skewness	-0.01825961396
Kurtosis	3.263530532311
Maximum	300.07
Upper Quartile	299.8975
Median	299.85
Lower Quartile	299.8025
Minimum	299.62
Range	0.45
Variance coeff.	0.000263498134
Sum	29985.24
Centile 5	299.721

Student's t test for paired samples

Data are from Bland (1996) and given in the columns "PEFR Before and PEFR After" of the "test" workbook.

Paired t test

For differences between PEFR Before and PEFR After:
Mean of differences = 56.111111
Standard deviation = 34.173983
Standard error = 11.391328

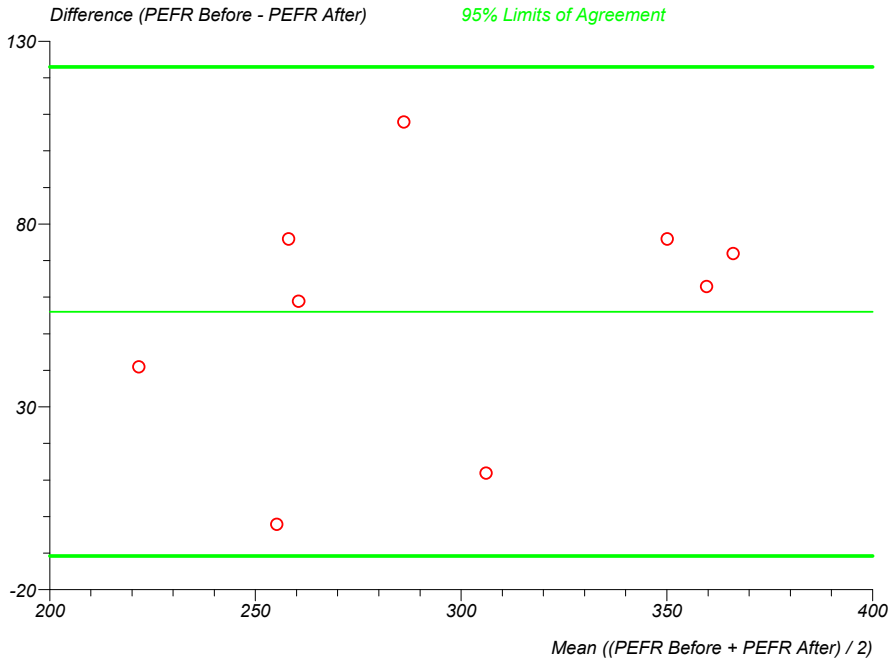
95% CI = 29.842662 to 82.37956

df = 8
t = 4.925774

One sided P = 0.0006
Two sided P = 0.0012

Two sample analysis of agreement

95% Limits of agreement = -10.868665 to 123.090887



Student's t test for a single sample

Data are from Altman (1991) and are provided in the "Energy intake" column of the "test" workbook.

Single sample t test

Sample name: Energy intake
Sample mean = 6753.636364
Population mean = 7725
Sample size n = 11
Sample sd = 1142.123222

95% confidence interval for mean difference = -1738.652457 to -204.074815

df = 10
t = -2.820754

One sided P = 0.0091
Two sided P = 0.0181

Student's t test for two independent samples

Data are from Armitage and Berry (1994) and are provided in the "High Protein" and "Low Protein" columns of the "test" workbook.

Unpaired t test

Mean of High Protein = 120
Mean of Low Protein = 101
Combined standard error = 10.045276

95% confidence interval for difference between means = -2.193679 to 40.193679

df = 17
t = 1.891436

One sided P = 0.0379
Two sided P = 0.0757

Two sided F test not significant - t test ok

F (variance ratio) test for two samples

Data are from Altman (1991) and are provided in the "Slight/no symptoms" and "Marked symptoms" columns of the "test" workbook.

Variance ratio/F test

Variable Name	DF (n-1)	Variance
Marked symptoms	6	1404.809524
Slight/no symptoms	8	202.277778

F = 6.944952

Upper side P = 0.0077
Two sided P = 0.0153

Normal distribution (z) test for a single sample

Data are given in the "SDI not conceived" column of the "test" workbook.

Normal distribution (z) test - single sample

Sample name:	SDI not conceived
Sample mean =	141.086207
Population mean =	145
Sample size n =	116
Sample sd =	14.616589
Population sd =	15

95% confidence interval for mean difference = -6.573692 to -1.253894

Standard normal deviate (z) = -2.810189
One sided P = 0.0025
Two sided P = 0.005

For lognormal data:
Geometric mean = 140.336754 (95% confidence interval = 114.485181 to 172.025798)

Normal distribution (z) test for two independent samples

Data are given in the "SDI not conceived" and "SDI conceived" columns of the "test" workbook.

Normal distribution (z) test - two independent samples

Sample name:	SDI conceived
Mean =	170.166667
Variance =	548.386179
Size =	42

Sample name:	SDI not conceived
Mean =	141.086207
Variance =	213.644678
Size =	116

Combined standard error = 3.859868

95% confidence interval for difference between means = 21.515258 to 36.645661

Standard normal deviate (Z) = 7.534056

One sided $P < 0.0001$
Two sided $P < 0.0001$

Reference range

Data are from Altman (1991) and are provided in the "IgM" column of the "test" workbook.

Reference range/interval

Sample name: IgM
Sample mean = 0.80302
Sample size n = 298
Sample sd = 0.469498

For normal data
95% reference interval = -0.117179 to 1.72322
95% confidence interval for lower range limit = -0.20828 to -0.026079
95% confidence interval for upper range limit = 1.632119 to 1.81432

For log-normal data
95% reference interval = 0.238102 to 2.031412
95% confidence interval for lower range limit = 0.214129 to 0.264758
95% confidence interval for upper range limit = 1.826887 to 2.258836

For any data
Quantile 0.025 = 0.2
95% confidence interval = 0.1 to 0.3

Quantile 0.975 = 2
95% confidence interval = 1.7 to 2.5

Poisson confidence interval

Data are from Rice (1995) and are provided in the "Fibres" column of the "test" workbook.

Poisson confidence interval

Sample name: Fibres
Size = 23
Mean = 24.913043
Approximate two sided 95% confidence interval = 22.914684 to 27.039011

Shapiro-Wilk W test

Data are from Shapiro and Wilk (1965) and are provided in the "Penicillin" column of the "test" workbook.

Shapiro-Wilk W test for non-normality

Sample name: Penicillin
Uncensored data = 30
Censored data = 0
Mean = -0.007033
Standard deviation = 0.0454
Squares about mean = 0.059774

W = 0.892184
P = 0.005437

Sample unlikely to be from a normal distribution

Mann-Whitney test

Data are from Conover (1999) and are provided in the "Town boys" and "Farm boys" columns of the "test" workbook.

Mann-Whitney U test

Observations (x) in Farm Boys = 12 median = 9.8 rank sum = 321
Observations (y) in Town Boys = 36 median = 7.75
U = 243 U' = 189

Exact probability (adjusted for ties):
Lower side P = 0.2645 (H_1: x tends to be less than y)
Upper side P = 0.7355 (H_1: x tends to be greater than y)
Two sided P = 0.529 (H_1: x tends to be distributed differently to y)

95.1% confidence interval for difference between medians or means:
K = 134 median difference = 0.8
CI = -2.3 to 4.4

Wilcoxon signed ranks test

Data are from Conover (1999) and provided in the "First Born" and "Second Born" columns of the "test" workbook.

Wilcoxon's signed ranks test

First Born vs. Second Born
Number of non-zero differences ranked = 11
Sum of ranks for positive differences = 41.5

Exact probability (adjusted for ties):
Lower side P = 0.7402 (H_1: differences tend to be less than zero)
Upper side P = 0.2598 (H_1: differences tend to be greater than zero)
Two sided P = 0.5195 (H_1: differences tend not to be zero)

95.8% confidence interval for difference between population medians:
K =14
CI = -2.5 to 6.5
Median difference = 1.5

Kendall's rank correlation

Data are from Armitage and Berry (1994) and are provided in the "Career" and "Psychology" columns of the "test" workbook.

Kendall's rank correlation

Career vs. Psychology
Observations per sample = 10
Concordant pairs = 34
Discordant pairs = 11
Tied pairs = 0
Kendall's score = 23 (standard error = 11.18034)
Gamma = 0.511111

Kendall's tau = 0.511111

Approximate 95% CI = 0.135203 to 0.887019

Approximate tests

Sample size too small for reliable inference from z, use exact test

z = 2.057183

Upper side P = 0.0198 (H_1: concordance)

Lower side P = 0.9802 (H_1: discordance)

Two sided P = 0.0397 (H_1: dependence)

z (continuity corrected) = 1.96774

Upper side P = 0.0245 (H_1: concordance)

Lower side P = 0.9755 (H_1: discordance)

Two sided P = 0.0491 (H_1: dependence)

Exact test

Upper side P = 0.0233 (H_1: concordance)

Lower side P = 0.9767 (H_1: discordance)

Two sided P = 0.0466 (H_1: dependence)

Spearman's rank correlation

Data are from Armitage and Berry (1994) and are provided in the "Career" and "Psychology" columns of the "test" workbook.

Spearman's rank correlation

Career vs. Psychology

Observations per sample = 10

Spearman's rank correlation coefficient (Rho) = 0.684848

95% CI for rho (Fisher's Z transformed) = 0.097085 to 0.918443

Upper side P = 0.0156 (H_1: positive correlation)

Lower side P = 0.9844 (H_1: negative correlation)

Two sided P = 0.0311 (H_1: any correlation)

Nonparametric linear regression

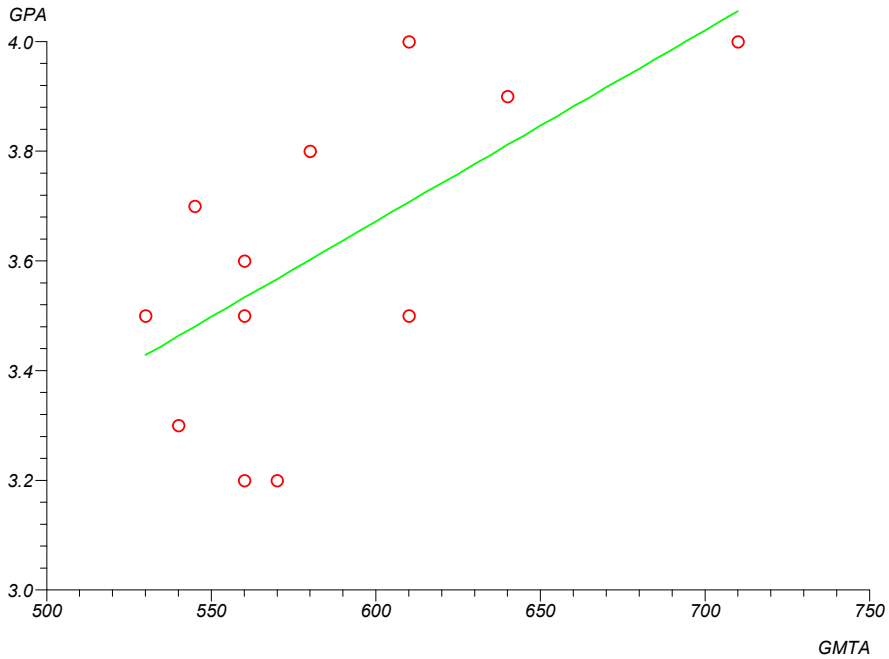
Data are from Conover (1999) and are provided in the "GPA" and "GMTA" columns of the "test" workbook.

Nonparametric linear regression

GPA vs. GMTA
Observations per sample = 12

Median slope (95% CI) = 0.003485 (0 to 0.0075)
Y-intercept = 1.581061

Kendall's rank correlation coefficient tau b = 0.439039
Two sided (on continuity corrected z) P = 0.0678 (H_1: any correlation)



Cuzick's test for trend

Data are from Cuzick (1985) and are provided in the "CMT 64", "CMT 167", "CMT 170", "CMT 175" and "CMT 181" columns of the "test" workbook.

Cuzick's trend test

Groups = 5
Observations = 45
Order: CMT 64, CMT 167, CMT 170, CMT 175, CMT 181

Ez = 3.022222 Var(z) = 1.93284
T = 3386.5 ET = 3128 Var(T) = 15003.666667
z = 2.110386

One sided P = 0.0174
Two sided P = 0.0348

Corrected for ties:
VarT = 14943.375253
z = 2.114639

One sided P = 0.0172
Two sided P = 0.0345

Smirnov two sample test

Data are from Conover (1999) and are provided in the "Xi" and "Yi" columns of the "test" workbook.

Two sample Smirnov test for distribution difference

X sample: Xi
Y sample: Yi

Two sided test:
D = 0.4
P = 0.2653

One sided test (suspecting Xi shifted left of Yi):
D = 0.4
P = 0.1326

One sided test (suspecting Xi shifted right of Yi):
D = 0.333333
P = 0.2432

Quantile confidence interval

Data are from Conover (1999) and are provided in the "Tubes" column of the "test" workbook.

Quantile confidence interval

Variable: Tubes
Sample size = 16
Quantile (0.75) = 66.35
Approximate 90% confidence interval = 63.3 to 73.3
Exact confidence level = 90.93936%

Kruskal-Wallis test

Data are from Conover (1999) and are provided in the "Method 1", "Method 2", "Method 3" and "Method 4" columns of the "test" workbook.

Kruskal-Wallis test

Variables: Method 1, Method 2, Method 3, Method 4

Groups = 4
df = 3
Total observations = 34
T = 25.464373
P < 0.0001

Adjusted for ties:
T = 25.628836
P < 0.0001

At least one of your sample populations tends to yield larger observations than at least one other.

Kruskal-Wallis: all pairwise comparisons (Dwass-Steel-Chritchlow-Fligner)
Critical q (range) = 3.63316

Method 1 vs. Method 2	not significant
(|2.97066| > 3.63316)	P = 0.1529
Method 1 vs. Method 3	not significant
(|3.385636| > 3.63316)	P = 0.0782
Method 1 vs. Method 4	significant
(|-4.91709| > 3.63316)	P = 0.0029
Method 2 vs. Method 3	significant
(|4.709793| > 3.63316)	P = 0.0048
Method 2 vs. Method 4	significant
(|-4.744037| > 3.63316)	P = 0.0044
Method 3 vs. Method 4	significant
(|4.59903| > 3.63316)	P = 0.0063

Kruskal-Wallis: all pairwise comparisons (Conover-Inman)

Critical t (30 df) = 2.042272

Method 1 vs. Method 2	significant
(6.533333 > 4.683291)	P = 0.0078
Method 1 vs. Method 3	significant
(7.738095 > 5.136713)	P = 0.0044
Method 1 vs. Method 4	significant
(17.020833 > 4.952835)	P < 0.0001
Method 2 vs. Method 3	significant
(14.271429 > 5.023091)	P < 0.0001
Method 2 vs. Method 4	significant
(10.4875 > 4.834893)	P = 0.0001
Method 3 vs. Method 4	significant
(24.758929 > 5.275301)	P < 0.0001

Squared ranks approximate equality of variance test

Chi-square = 6.006228 df = 3 P = 0.1113

Friedman test

Data are from Conover (1999) and are provided in the "Grass 1", "Grass 2", "Grass 3" and "Grass 4" columns of the "test" workbook.

Friedman test

Variables: Grass 1, Grass 2, Grass 3, Grass 4

Mean rank: 3.17, 1.96, 2.04, 2.83

Treatment average sum of squares of ranks = 356.5

Number of blocks = 12

T_1 (chi-square) = 8.097345

df = 3

After Iman & Davenport (1980):
T_2 (F) = 3.192198
P = 0.0362

At least one of your sample populations tends to yield larger observations than at least one other sample population.

Friedman: all pairwise comparisons (Conover)
Critical t (33 df) = 11.481678

Grass 1 vs. Grass 2	significant
(|14.5| > 11.481678)	P = 0.0149
Grass 1 vs. Grass 3	significant
(|13.5| > 11.481678)	P = 0.0226
Grass 1 vs. Grass 4	not significant
(|4| > 11.481678)	P = 0.4834
Grass 2 vs. Grass 3	not significant
(|-1| > 11.481678)	P = 0.8604
Grass 2 vs. Grass 4	not significant
(|-10.5| > 11.481678)	P = 0.0717
Grass 3 vs. Grass 4	not significant
(|-9.5| > 11.481678)	P = 0.1017

Squared ranks approximate equality of variance test
Chi-square = 6.169426 df = 3 P = 0.1037

Chi-square goodness of fit test

Data are from Conover (1999) and are provided in the "Digits observed" column of the "test" workbook.

Chi-square goodness of fit

Sample: Digits observed:

N = 300

Value	Observed frequency	Expected frequency
1	22	30
2	28	30
3	41	30
4	35	30
5	19	30
6	25	30
7	25	30
8	40	30
9	30	30
10	35	30

Chi-square = 17 df = 9

P = 0.0487

One way analysis of variance

The first set of data is from Kleinbaum et al. (1998) and is provided in the "Substance 1", " Substance 2", " Substance 3" and " Substance 4" columns of the "test" workbook. The second set of data is from and is provided in the "Instrument 1", "Instrument 2", " Instrument 3", " Instrument 4", and " Instrument 5" columns of the "test" workbook. Results for the second data set are displayed with extended decimal precision as they are intended for testing precision of analysis of variance algorithms and they form part of the Statistical Reference Data Set (www.nist.gov.itl/div898/strd) from the American National Institute of Standards and Technology.

One way analysis of variance

Variables: Substance 1, Substance 2, Substance 3, Substance 4

Source of Variation	Sum Squares	DF	Mean Square
Between Groups	249.875	3	83.291667
Within Groups	350.9	36	9.747222
Corrected Total	600.775	39	

F (variance ratio) = 8.54517 P = 0.0002

Tukey multiple comparisons

Critical value (Studentized range) = 3.808798, |q*| = 2.693292
Pooled standard deviation = 3.122054

Comparison	Mean difference L (95% CI)	|L/SE(L)|	
Substance 1 vs. Substance 4	6.3 (2.539649 to 10.060351)	6.381167	P = 0.0004
Substance 1 vs. Substance 3	5.9 (2.139649 to 9.660351)	5.976014	P = 0.0009
Substance 1 vs. Substance 2	3.7 (-0.060351 to 7.460351)	3.74767	P = 0.0552 stop
Substance 2 vs. Substance 4	2.6 (-1.160351 to 6.360351)	2.633498	P = 0.2621
Substance 2 vs. Substance 3	2.2 (-1.560351 to 5.960351)	2.228344	P = 0.4049
Substance 3 vs. Substance 4	0.4 (-3.360351 to 4.160351)	0.405153	P = 0.9917

Scheffé multiple comparisons

Critical value = 2.93237

Comparison	Mean difference L (95% CI)	|L/SE(L)|	
Substance 1 vs. Substance 4	6.3 (2.205751 to 10.394249)	4.512167	P = 0.001
Substance 1 vs. Substance 3	5.9 (1.805751 to 9.994249)	4.22568	P = 0.0021
Substance 1 vs. Substance 2	3.7 (-0.394249 to 7.794249)	2.650003	P = 0.0896 stop
Substance 2 vs. Substance 4	2.6 (-1.494249 to 6.694249)	1.862164	P = 0.34
Substance 2 vs. Substance 3	2.2 (-1.894249 to 6.294249)	1.575677	P = 0.4874
Substance 3 vs. Substance 4	0.4 (-3.694249 to 4.494249)	0.286487	P = 0.9938

Newman-Keuls multiple comparisons

Comparison	Mean difference L	Separation	|L/SE(L)|	
Substance 1 vs. Substance 4	6.3	4	6.381167	P = 0.0004
Substance 1 vs. Substance 3	5.9	3	5.976014	P = 0.0004
Substance 1 vs. Substance 2	3.7	2	3.74767	P = 0.0119
Substance 2 vs. Substance 4	2.6	3	2.633498	P = 0.1644 stop
Substance 2 vs. Substance 3	2.2	2	2.228344	P = 0.1238
Substance 3 vs. Substance 4	0.4	2	0.405153	P = 0.7761

Dunnett multiple comparisons with a control

Critical value (|d|) = 2.452195
Pooled standard deviation = 3.122054

Control (n) = Substance 1 (10)

Comparison (n)	Mean difference (95% CI)	
Substance 4 (10)	-6.3 (-9.723816 to -2.876184)	P = 0.0002
Substance 3 (10)	-5.9 (-9.323816 to -2.476184)	P = 0.0005
Substance 2 (10)	-3.7 (-7.123816 to -0.276184)	P = 0.0316

Approximate equality of variance tests

Levene's (W50) F = 0.511364 (df = 3, 36) P = 0.677

Bartlett's chi-square = 0.343929 df = 3 P = 0.9516

One way analysis of variance

Variables: Instrument 1, Instrument 2, Instrument 3, Instrument 4, Instrument 5

Source of Variation	Sum Squares	DF	Mean Square
Between Groups	0.0511462616	4	0.0127865654
Within Groups	0.21663656	20	0.010831828
Corrected Total	0.2677828216	24	

F (variance ratio) = 1.180462374402 P = 0.3494

Two way randomized blocks analysis of variance

Data are from Armitage and Berry (1994) and are provided in the "Treatment 1", "Treatment 2", "Treatment 3" and "Treatment 4" columns of the "test" workbook.

Two way randomized block analysis of variance

Variables: Treatment 1, Treatment 2, Treatment 3, Treatment 4

Source of Variation	Sum Squares	DF	Mean Square
Between blocks (rows)	78.98875	7	11.284107
Between treatments (columns)	13.01625	3	4.33875
Residual (error)	13.77375	21	0.655893
Corrected total	105.77875	31	

F (VR between blocks) = 17.204193 P < 0.0001

F (VR between treatments) = 6.615029 P = 0.0025

Two way replicate randomized blocks analysis of variance

Data are from Armitage and Berry (1994) and are provided in the "T1(rep 1)", "T2(rep 1)", "T3(rep 1)", "T1(rep 2)", "T2(rep 2)", "T3(rep 2)", "T1(rep 3)", "T2(rep 3)" and "T3(rep 3)" columns of the "test" workbook.

Two way randomized block analysis of variance with repeated observations

Variables: (T1 (rep 1), T2 (rep 1), T3 (rep 1)) (T1 (rep 2), T2 (rep 2), T3 (rep 2)) (T1 (rep 3), T2 (rep 3), T3 (rep 3))

Source of Variation	Sum Squares	DF	Mean Square
Blocks (rows)	9.26	2	4.63
Treatments (columns)	11.78	2	5.89
Interaction	0.74	4	0.185
Residual (error)	1.32	18	0.073333
Corrected total	23.1	26	

F (VR blocks) = 63.136364 $P < 0.0001$

F (VR treatments) = 80.318182 $P < 0.0001$

F (VR interaction) = 2.522727 $P = 0.0771$

Nested random analysis of variance

Data are from Snedecor and Cochran (1989) and are provided in the "P1L1", "P1L2", "P1L3", "P2L1", "P2L2", "P2L3", "P3L1", "P3L2", "P3L3", "P4L1", "P4L2", "P4L3" columns of the "test" workbook.

Fully nested/hierarchical random analysis of variance

Variables: (P1L1, P1L2, P1L3) (P2L1, P2L2, P2L3) (P3L1, P3L2, P3L3) (P4L1, P4L2, P4L3)

Source of Variation	Sum Squares	DF	Mean Square
Between Groups	7.560346	3	2.520115
Between Subgroups within Groups	2.6302	8	0.328775
Residual	0.07985	12	0.006654
Total	10.270396	23	

F (VR between groups) = 378.727406 P < 0.0001

F (using group/subgroup msqr) = 7.665167 P = 0.0097

F (VR between subgroups within groups) = 49.408892 P < 0.0001

Group 1 mean = 3.175 (n = 6)

Group 2 mean = 2.178333 (n = 6)

Group 3 mean = 2.951667 (n = 6)

Group 4 mean = 3.743333 (n = 6)

Grand mean = 3.012083 (n = 24)

Group 1	(subgroup 1)	P1L1	mean = 3.185
Group 1	(subgroup 2)	P1L2	mean = 3.5
Group 1	(subgroup 3)	P1L3	mean = 2.84
Group 2	(subgroup 1)	P2L1	mean = 2.45
Group 2	(subgroup 2)	P2L2	mean = 1.895
Group 2	(subgroup 3)	P2L3	mean = 2.19
Group 3	(subgroup 1)	P3L1	mean = 2.715
Group 3	(subgroup 2)	P3L2	mean = 3.59
Group 3	(subgroup 3)	P3L3	mean = 2.55
Group 4	(subgroup 1)	P4L1	mean = 3.825
Group 4	(subgroup 2)	P4L2	mean = 4.095
Group 4	(subgroup 3)	P4L3	mean = 3.31

Latin square

Data are from Armitage and Berry (1994) and are provided in the "Rabbit 1", "Rabbit 2", "Rabbit 3", "Rabbit 4", "Rabbit 5", "Rabbit 6" and "Trtmnt Total" columns of the "test" workbook.

Latin square test

Variables: Rabbit 1, Rabbit 2, Rabbit 3, Rabbit 4, Rabbit 5, Rabbit 6, Trtmnt Total

Source of Variation	Sum Squares	DF	Mean Square
Rows	3.833333	5	0.766667
Columns	12.833333	5	2.566667
Treatments	0.563333	5	0.112667
Residual	13.13	20	0.6565
Total	30.36	35	

F (rows) = 1.167809 P = 0.3592

F (columns) = 3.909622 P = 0.0124

F (treatments) = 0.171617 P = 0.9701

Crossover

Data are from Armitage and Berry (1994) and are provided in the "Drug 1", "Placebo 1", "Drug 2" and "Placebo 2" columns of the "test" workbook.

Crossover tests

	Period 1	Period 2	Difference
Group 1	8.117647	5.294118	2.823529
Group 2	7.666667	8.916667	-1.25

Test for relative effectiveness of drug / placebo:
combined diff = 2.172414 SE = 0.61602
t = 3.526533 DF = 28 P = 0.0015

Test for treatment effect:
diff 1 - diff 2 = 4.073529 SE = 1.2372
effect magnitude = 2.036765 95% confidence interval = 0.767502 to 3.306027
t = 3.292539 DF = 27 P = 0.0028

Test for period effect:
diff 1 + diff 2 = 1.573529 SE = 1.2372
t = 1.271847 DF = 27 P = 0.2143

Test for treatment-period interaction:
sum 1 - sum 2 = -3.171569 SE = 2.440281
t = -1.299673 DF = 27 P = 0.2047

Agreement analysis

Data are from Bland and Altman (1996a) and are provided in the columns marked "1st", "2nd", "3rd" and "4th" in the "test" workbook.

Agreement

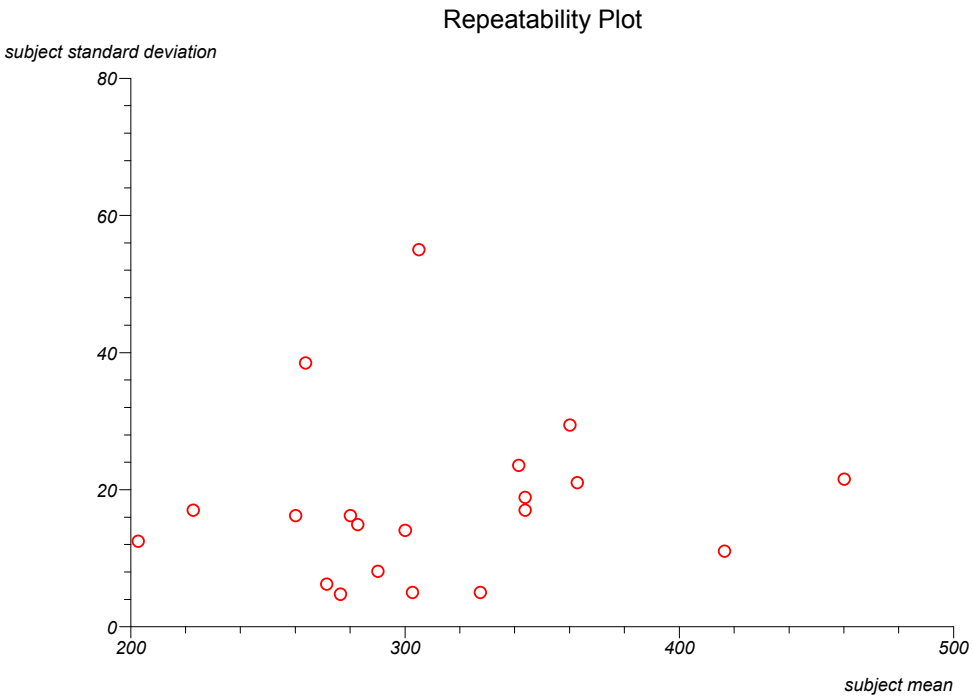
Variables: 1st, 2nd, 3rd, 4th

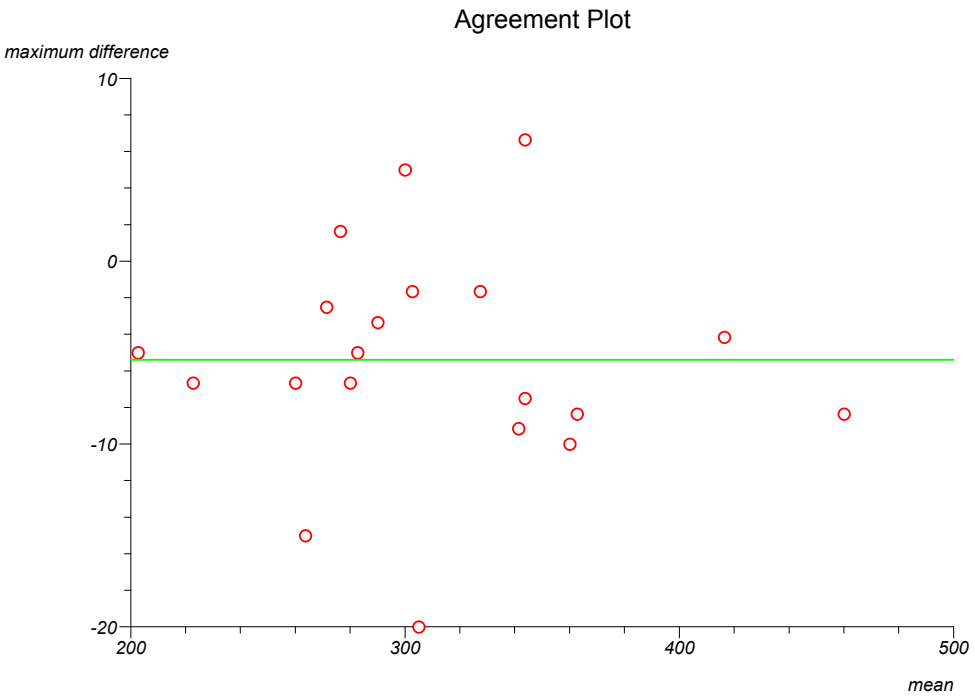
Intra-class correlation coefficient (one way random effects) = 0.882276

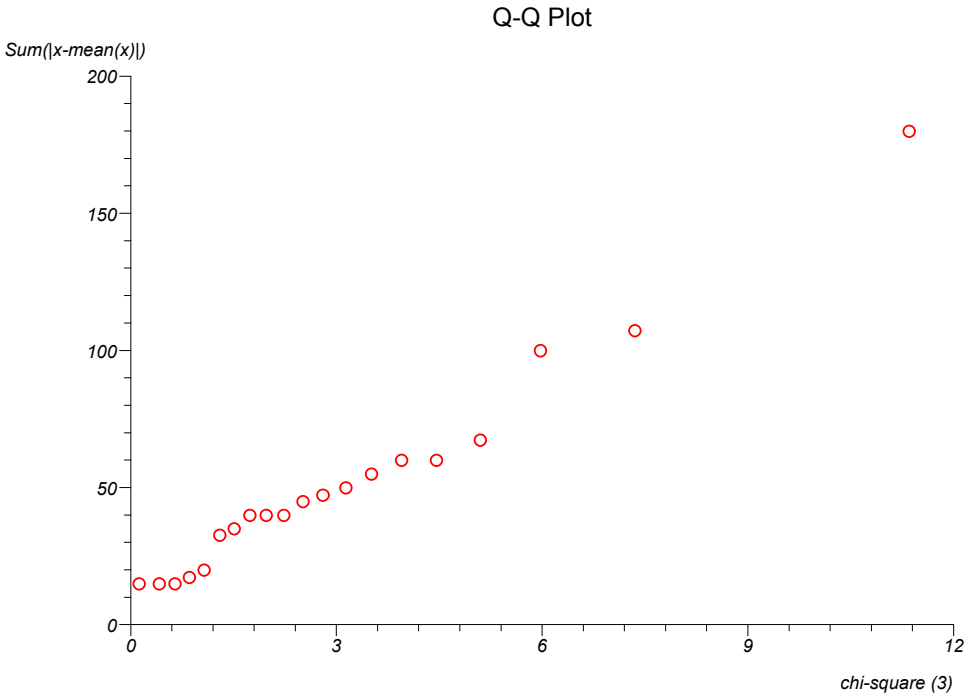
Estimated within-subjects standard deviation = 21.459749

For within-subjects sd vs. mean, Kendall's tau b = 0.164457 two sided P = 0.3304

Repeatability (for alpha = 0.05) = 59.482297




Agreement Plot
maximum difference
10
0
-10
-20
200
300
400
500
mean


Q-Q Plot
Sum(|x-mean(x)|)
200
150
100
50
0
0
3
6
9
12
chi-square (3)

Simple linear regression

Data are from Armitage and Berry (1994) and are provided in the "Birth Weight" and "% Increase" columns of the "test" workbook.

Simple linear regression

Equation: % Increase = -0.86433 Birth Weight + 167.870079

Standard Error of slope = 0.175684
95% CI for population value of slope = -1.223125 to -0.505535

Correlation coefficient (r) = -0.668236 (r^2 = 0.446539)

95% CI for r (Fisher's z transformed) = -0.824754 to -0.416618

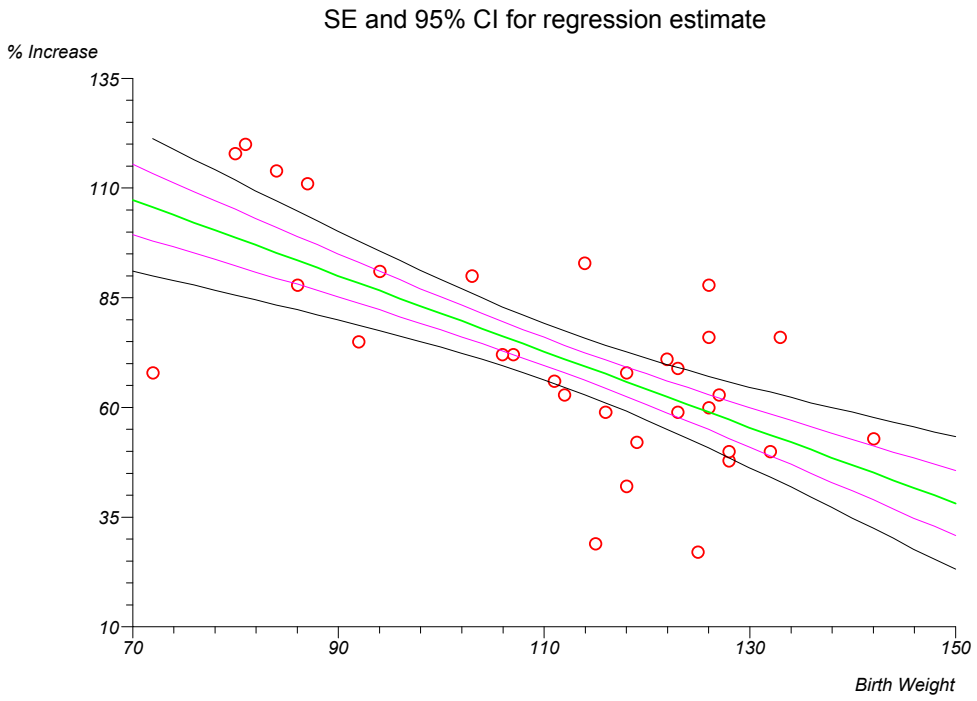
t with 30 DF = -4.919791
Two sided P < 0.0001
Correlation coefficient is significantly different from zero

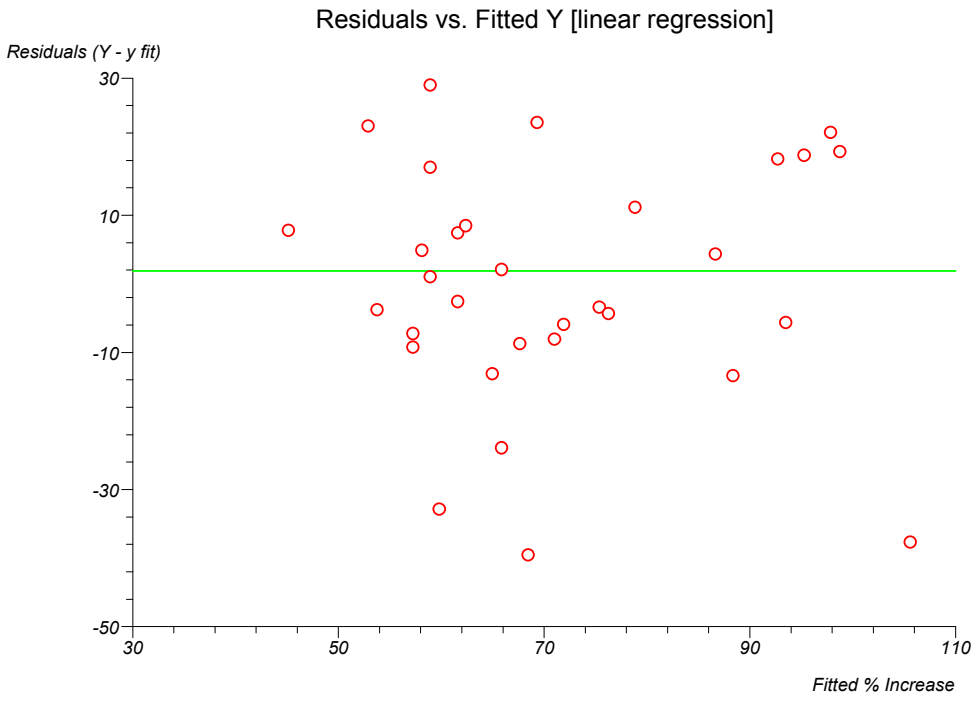
Simple regression - analysis

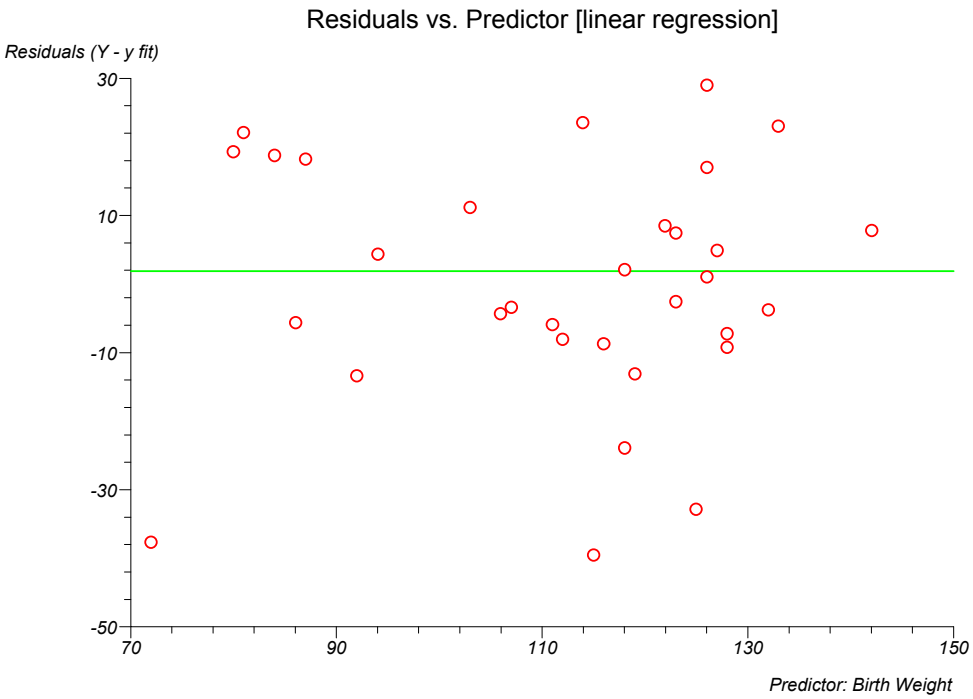
Source of variation	Sum Squares	DF	Mean Square
Regression	7666.387163	1	7666.387163
Residual	9502.081587	30	316.736053
Total, corrected	17168.46875	31	

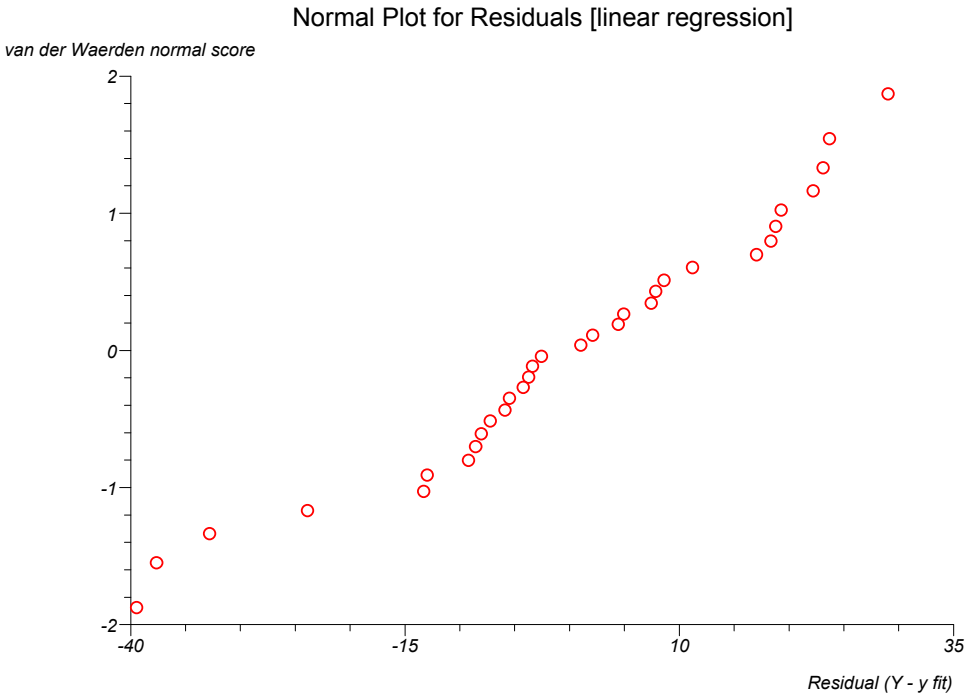
F = 24.20434 P < 0.0001

R square = 0.446539
Root MSE = 17.79708


SE and 95% CI for regression estimate
% Increase
135
110
85
60
35
10
70
90
110
130
150
Birth Weight


Residuals vs. Fitted Y [linear regression]
Residuals (Y - y fit)
30
10
-10
-30
-50
30
50
70
90
110
Fitted % Increase


Residuals vs. Predictor [linear regression]
Residuals (Y - y fit)
30
10
-10
-30
-50
70
90
110
130
150
Predictor: Birth Weight


Normal Plot for Residuals [linear regression]
van der Waerden normal score
2
1
0
-1
-2
-40
-15
10
35
Residual (Y - y fit)

Multiple/general linear regression

Data are from Armitage and Berry (1994) and from Longley (1967), and are provided in the "YY", "X1", "X2", "y (Longley)", "x1 (Longley)", "x2 (Longley)", "x3 (Longley)", "x4 (Longley)", "x5 (Longley)" and "x6 (Longley)" columns of the "test" workbook. The Longley data are also provided for this purpose by the American National Institute of Standards and Technology at www.nist.gov/itl/div898/strd.

Multiple linear regression

Intercept	b0 = 23.010668	t = 1.258453	P = 0.2141
x1	b1 = 23.638558	t = 3.45194	P = 0.0011
x2	b2 = -0.714675	t = -2.371006	P = 0.0216

yy = 23.010668 +23.638558 x1 -0.714675 x2

Analysis of variance from regression

Source of variation	Sum Squares	DF	Mean Square
Regression	2783.220444	2	1391.610222
Residual	11007.949367	50	220.158987
Total (corrected)	13791.169811	52	

Root MSE = 14.837755

F = 6.320933 P = 0.0036

Multiple correlation coefficient (R) = 0.449235

R^2 = 20.181177%

Ra^2 = 16.988424%

Durbin-Watson test statistic = 1.888528

Multiple linear regression - XXi (hat) matrix

1.518617	-0.132786	-0.018619
-0.132786	0.213	-0.004396
-0.018619	-0.004396	0.000413

Multiple linear regression - variance-covariance matrix

334.337183	-29.234028	-4.099146
-29.234028	46.893801	-0.967742
-4.099146	-0.967742	0.090856

Multiple linear regression - Parameter detail

	COEFFICIENT	STANDARD DEVIATION
constant	23.010668	18.284889
x1	23.638558	6.847905
x2	-0.714675	0.301423

Multiple linear regression - influential data

Index	Y	Fitted Y	Std. Dev. Y fit	Residual
1	7	29.265262	2.77435	-22.265262
2	10	28.633361	4.396969	-18.633361
3	18	13.630705	3.392506	4.369295
4	4	11.530825	3.790998	-7.530825
5	10	22.393526	2.159429	-12.393526
6	13	13.399837	3.375359	-0.399837
7	21	20.633981	6.171899	0.366019
8	12	28.545069	2.756796	-16.545069
9	9	23.39425	2.837475	-14.39425
10	65	25.680851	2.998838	39.319149
11	20	22.63543	2.087495	-2.63543
12	31	26.00001	3.041507	4.99999
13	23	7.308994	5.413088	15.691006
14	22	23.580972	2.11719	-1.580972
15	13	13.883645	3.211483	-0.883645
16	9	24.834637	3.644819	-15.834637
17	50	20.040706	2.170843	29.959294

18	12	17.440465	2.524752	-5.440465
19	11	24.290129	2.162085	-13.290129
20	8	21.503166	2.942601	-13.503166
21	26	15.543862	2.864326	10.456138
22	16	15.807839	3.138961	0.192161
23	23	27.379696	2.426817	-4.379696
24	7	24.977213	2.736885	-17.977213
25	11	20.563141	3.231827	-9.563141
26	8	12.228946	3.581038	-4.228946
27	14	29.715961	3.205115	-15.715961
28	39	34.663503	4.747572	4.336497
29	28	27.902131	3.523884	0.097869
30	12	29.407839	5.749988	-17.407839
31	60	28.786973	2.729682	31.213027
32	10	4.109063	6.641434	5.890937
33	60	36.345794	4.566751	23.654206
34	22	19.820875	2.348754	2.179125
35	21	22.915961	2.669797	-1.915961
36	14	16.049742	3.221507	-2.049742
37	4	25.664297	3.605758	-21.664297
38	27	19.133791	3.017109	7.866209
39	26	33.080541	3.703593	-7.080541
40	28	21.030394	2.984058	6.969606
41	15	20.892439	5.164145	-5.892439
42	8	16.500441	2.73245	-8.500441
43	46	30.436154	3.142523	15.563846
44	24	29.270781	2.75301	-5.270781
45	12	19.084128	2.327277	-7.084128
46	25	18.594802	2.6631	6.405198
47	45	25.026877	2.269734	19.973123
48	72	38.539482	5.197819	33.460518
49	25	15.532826	2.910596	9.467174
50	28	29.485093	3.039633	-1.485093
51	10	17.715478	3.234597	-7.715478
52	25	36.203217	5.409996	-11.203217
53	44	21.9649	2.547996	22.0351

Index		Studentized		Leverage	Cook's Distance
1		-1.527521		0.034961	0.028177
2		-1.314866		0.087815	0.055479
3		0.302484		0.052276	0.001682
4		-0.524969		0.065279	0.006416
5		-0.844258		0.021181	0.005141
6		-0.027673		0.051749	0.000014
7		0.027126	*	0.173022	0.000051
8		-1.134825		0.03452	0.015348
9		-0.98835		0.03657	0.01236
10	*	2.705778		0.040848	0.103931
11		-0.179401		0.019793	0.000217
12		0.344288		0.042019	0.001733
13		1.135785		0.133093	0.066016
14		-0.107652		0.02036	0.00008
15		-0.061		0.046846	0.000061
16		-1.100918		0.060341	0.025944
17	*	2.041089		0.021405	0.030375
18		-0.37209		0.028954	0.001376
19		-0.90536		0.021233	0.005927
20		-0.928497		0.03933	0.011765
21		0.718207		0.037266	0.006655
22		0.013251		0.044754	0.000003
23		-0.299201		0.026751	0.00082
24		-1.232738		0.034023	0.017841
25		-0.660369		0.047442	0.00724
26		-0.293694		0.058248	0.001778
27		-1.084798		0.046661	0.019199
28		0.308478		0.102378	0.003618
29		0.00679		0.056404	9.19E-07
30		-1.272658		0.150175	0.095405
31	*	2.14015		0.033844	0.053482
32		0.443983	*	0.200349	0.016463
33		1.675524		0.094728	0.097922
34		0.148739		0.025058	0.00019
35		-0.13127		0.032376	0.000192
36		-0.14152		0.047139	0.00033
37		-1.5052		0.059055	0.047398
38		0.54146		0.041347	0.004215

39		-0.492796	0.062303	0.005378
40		0.479518	0.040446	0.003231
41		-0.423609	0.121132	0.008244
42		-0.582861	0.033913	0.003975
43		1.073283	0.044856	0.018033
44		-0.361505	0.034425	0.001553
45		-0.483423	0.024601	0.001965
46		0.438808	0.032214	0.002136
47		1.362133	0.0234	0.014819
48	*	2.407657	0.122717	0.270293
49		0.650688	0.038479	0.005648
50		-0.102258	0.041967	0.000153
51		-0.532804	0.047523	0.004721
52		-0.810868	0.132941	0.033604
53		1.507463	0.029489	0.023016

Index		Jackknife		DFIT
1		-1.548737		-0.29478
2		-1.324756		-0.411036
3		0.299718		0.070392
4		-0.521131		-0.137718
5		-0.841795		-0.12383
6		-0.027395		-0.0064
7		0.026854		0.012283
8		-1.138172		-0.215215
9		-0.988117		-0.192514
10	*	2.899241	*	0.598309
11		-0.177655		-0.025245
12		0.341233		0.071465
13		1.139161		0.44635
14		-0.106583		-0.015365
15		-0.060389		-0.013388
16		-1.103307		-0.279588
17	*	2.110407		0.312122
18		-0.368861		-0.063693
19		-0.903699		-0.133103
20		-0.927193		-0.187606
21		0.714685		0.14061
22		0.013118		0.002839

23		-0.29646		-0.04915
24		-1.239327		-0.23259
25		-0.656601		-0.146533
26		-0.290994		-0.07237
27		-1.08676		-0.240428
28		0.305669		0.10323
29		0.006722		0.001643
30		-1.280783	*	-0.538405
31	*	2.222899		0.416045
32		0.440389		0.220435
33		1.707307	*	0.552282
34		0.147277		0.023611
35		-0.129973		-0.023774
36		-0.140125		-0.031167
37		-1.525024		-0.382052
38		0.537597		0.111647
39		-0.489032		-0.126055
40		0.475794		0.097684
41		-0.420106		-0.155965
42		-0.578973		-0.108476
43		1.074951		0.232951
44		-0.35834		-0.067662
45		-0.479686		-0.076181
46		0.435237		0.079406
47		1.37418		0.212712
48	*	2.53493	*	0.948089
49		0.646893		0.12941
50		-0.10124		-0.021189
51		-0.528953		-0.118152
52		-0.808049		-0.316404
53		1.527425		0.26625

Critical levels for unusual observations (marked *)

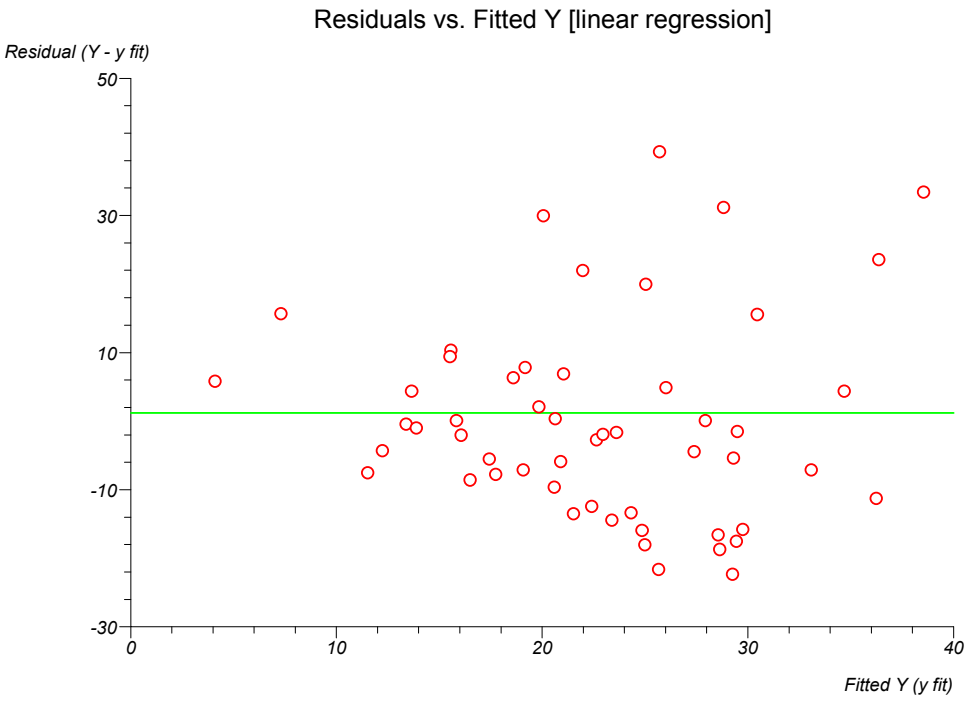
Leverage > 0.169811 [min(3p/n,.99)]

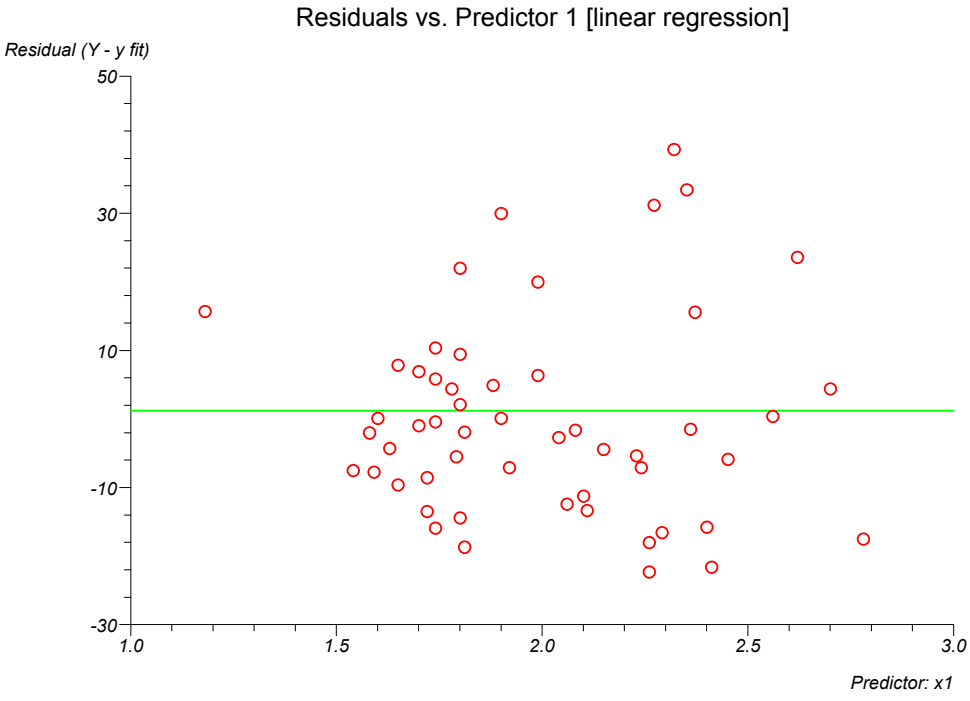
Cook's Distance > 2.790008 [f(α,p,n-p]

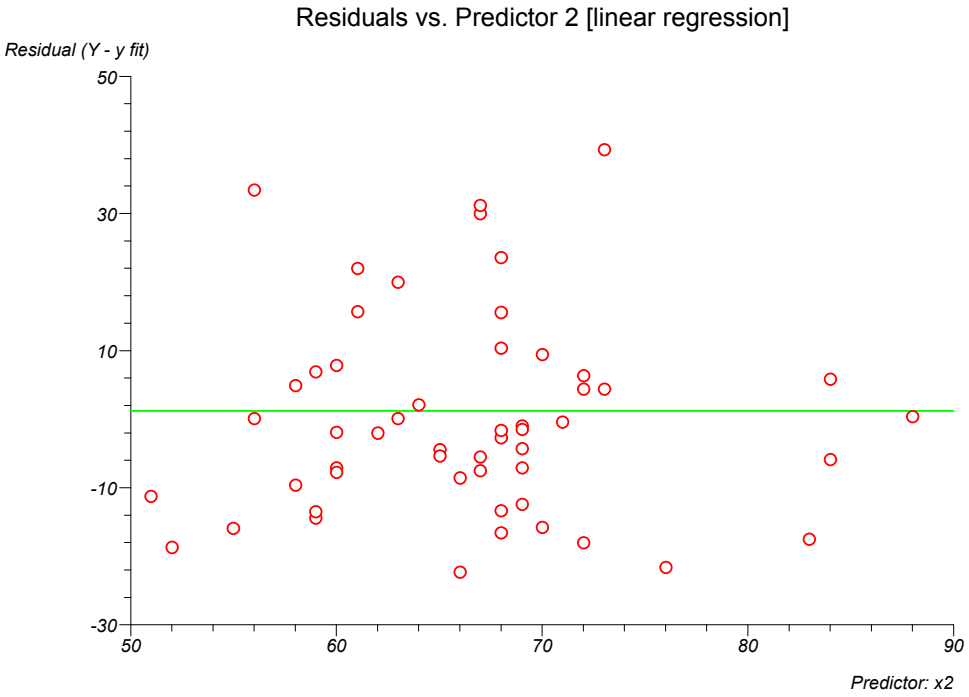
Studentized residual > ± 2.008559 [t(α,n-p)]

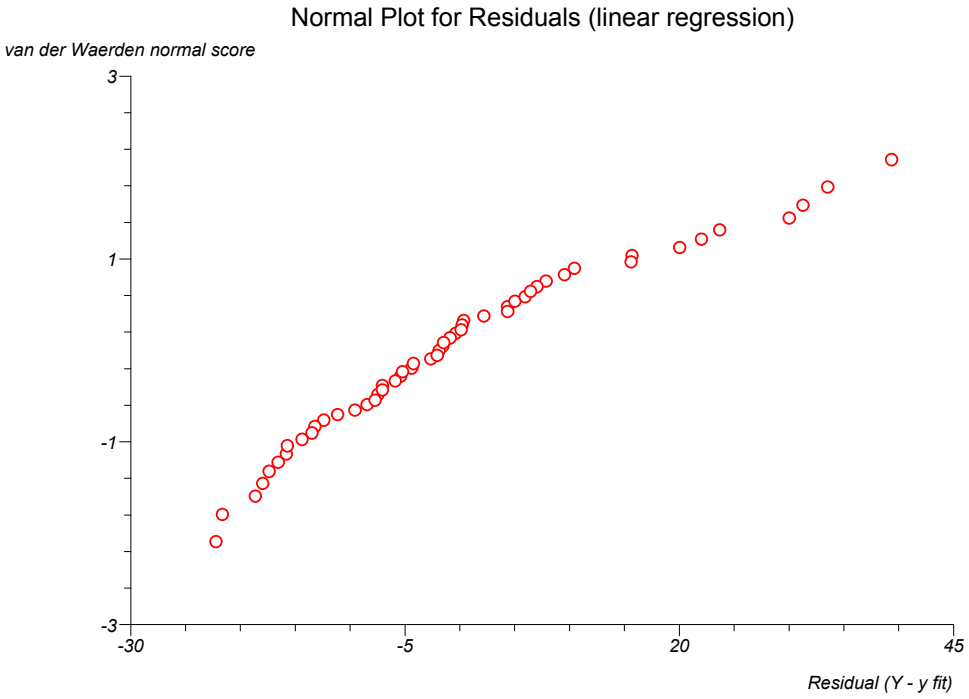
Jackknife residual > ± 2.009575 [t(α,n-p-1)]

DFIT > ± 0.475831 [(2*sqr(p/n)]


Residuals vs. Fitted Y [linear regression]
Residual (Y - y fit)
50
30
10
-10
-30
0
10
20
30
40
Fitted Y (y fit)


Residuals vs. Predictor 1 [linear regression]
Residual (Y - y fit)
50
30
10
-10
-30
1.0
1.5
2.0
2.5
3.0
Predictor: x1


Residuals vs. Predictor 2 [linear regression]
Residual (Y - y fit)
50
30
10
-10
-30
50
60
70
80
90
Predictor: x2


Normal Plot for Residuals (linear regression)
van der Waerden normal score
3
1
-1
-3
-30
-5
20
45
Residual (Y - y fit)

Multiple linear regression

Intercept	b0 = -3482258.634596	t = -3.910803	P = 0.0036
x1 (Longley)	b1 = 15.061872	t = 0.177376	P = 0.8631
x2 (Longley)	b2 = -0.035819	t = -1.069516	P = 0.3127
x3 (Longley)	b3 = -2.02023	t = -4.136427	P = 0.0025
x4 (Longley)	b4 = -1.033227	t = -4.821985	P = 0.0009
x5 (Longley)	b5 = -0.051104	t = -0.226051	P = 0.8262
x6 (Longley)	b6 = 1829.151465	t = 4.01589	P = 0.003

y (Longley) = -3482258.634596 +15.061872 x1 (Longley) -0.035819 x2 (Longley) -2.02023 x3 (Longley) -1.033227 x4 (Longley) -0.051104 x5 (Longley) +1829.151465 x6 (Longley)

Multiple linear regression - parameter detail

	Coefficient	Standard Deviation
constant	-3482258.634596	890420.383607
x1 (Longley)	15.061872	84.914926
x2 (Longley)	-0.035819	0.033491
x3 (Longley)	-2.02023	0.4884
x4 (Longley)	-1.033227	0.214274
x5 (Longley)	-0.051104	0.226073
x6 (Longley)	1829.151465	455.478499

Analysis of variance from regression

Source of variation	Sum Squares	DF	Mean Square
Regression	184172401.944494	6	30695400.324082
Residual	836424.055506	9	92936.006167
Total (corrected)	185008826	15	

Root MSE = 304.854074

F = 330.285339 P < 0.0001

Multiple correlation coefficient (R) = 0.997737

R^2 = 99.5479%

Ra^2 = 99.246501%

Durbin-Watson test statistic = 2.403375

Grouped regression - linearity

Data are from Armitage and Berry (1994) and are provided in the "Log_Dose_Std", "BD1_Std", "BD2_Std" and "BD3_Std" columns of the "test" workbook.

Linearity with replicates of Y

Source of variation	SSq	DF	MSq	VR	
Due to regression	14.088629	1	14.088629	9.450512	P = 0.0047
Deviation of x means	2.903415	1	2.903415	1.947581	P = 0.1738
Within x residual	41.741827	28	1.49078		
Total	58.733871	30			

Regression slope is significant
Assumption of linearity supported

Grouped regression - covariance

Data are from Armitage and Berry (1994) and are provided in the "Age exposed >10y", "VC exposed >10y", "Age exposed <10y", "VC exposed <10y", "Age not exposed" and "VC not exposed" columns of the "test" workbook.

Grouped linear regression

Source of variation	SSq	DF	MSq	VR	
Common slope	14.858947	1	14.858947	42.091481	P < 0.0001
Between slopes	2.499458	2	1.249729	3.540153	P = 0.0338
Separate residuals	27.535213	78	0.353016		
Within groups	44.893618	81			

Common slope is significant
Difference between slopes is significant

Slope comparisons:
slope 1 (Age exposed > 10y) vs. slope 2 (Age exposed < 10y) = -0.085111 vs. -0.046532
Difference (95% CI) = 0.038579 (-0.007753 to 0.084911)
t = -1.657691 P = 0.1014

slope 1 (Age exposed > 10y) vs. slope 3 (Age not exposed) = -0.085111 vs. -0.030613
Difference (95% CI) = 0.054498 (0.012552 to 0.096445)
t = -2.586561 P = 0.0116
slope 2 (Age exposed < 10y) vs. slope 3 (Age not exposed) = -0.046532 vs. -0.030613
Difference (95% CI) = 0.015919 (-0.013041 to 0.04488)
t = -1.094348 P = 0.2772

Covariance analysis

Uncorrected:

Source of variation	YY	xY	xx	DF
Between groups	2.747338	-57.389719	1254.686147	2
Within	44.893618	-373.573377	9392.123377	81
Total	47.640956	-430.963095	10646.809524	83

Corrected:

Source of variation	SSq	DF	MSq	VR
Between groups	0.161699	2	0.080849	0.215349
Within	30.034671	80	0.375433	
Total	30.196369	82		

P = 0.8067

Corrected Y means ± SE for baseline mean predictor of 40.547619:
Y'= 4.315193 ± 0.186202
Y'= 4.36193 ± 0.117104
Y'= 4.432128 ± 0.092494

Line separations (common slope =-0.039775):
Lines not parallel
line 1 (Age exposed > 10y) vs. line 2 (Age exposed < 10y) vertical separation = -0.046737
95% CI = -0.493578 to 0.400103
t = -0.208151 (80 df) P = 0.8356
line 1 (Age exposed > 10y) vs. line 3 (Age not exposed) vertical separation = -0.116935
95% CI = -0.533328 to 0.299458
t = -0.558866 (80 df) P = 0.5778
line 2 (Age exposed < 10y) vs. line 3 (Age not exposed) vertical separation = -0.070198
95% CI = -0.366058 to 0.225663
t = -0.472174 (80 df) P = 0.6381

Principal components analysis

Data are from Johnson and Wichern (1998) and are provided in the "Pop", "School", "Employ", "Health" and "Home" columns of the "test" workbook.

Principal components

Component	Eigenvalue (SVD)	Proportion	Cumulative
1	3.028896	60.58%	60.58%
2	1.291138	25.82%	86.4%
3	0.572456	11.45%	97.85%
4	0.095398	1.91%	99.76%
5	0.012112	0.24%	100%

Variable	PC1	PC2	PC3	PC4	PC5
Pop	-0.558359	0.131393	0.007946	-0.550553	-0.606465
School	-0.313283	0.628873	-0.549031	0.452654	0.006565
Employ	-0.568258	0.004262	0.11728	-0.268116	0.769041
Health	-0.486625	-0.309561	0.454924	0.647982	-0.201326
Home	0.174266	0.701006	0.691225	-0.015107	0.014203

Row	PCS1	PCS2	PCS3	PCS4	PCS5
1	-0.598312	0.619445	0.445956	-0.422185	-0.2063
2	2.363046	-0.139106	-0.110269	0.177784	-0.143831
3	1.415717	-1.224911	-0.616336	-0.250236	0.020187
4	0.608641	1.398234	-0.421344	-0.06269	0.043652
5	-0.659299	0.04537	-0.356157	0.1859	0.154443
6	-3.28113	0.384768	0.247039	0.128708	0.034615
7	1.314041	-0.666078	-0.645174	-0.134602	-0.003156
8	-1.946234	0.911026	-1.654572	0.343382	-0.1037
9	-2.338702	-1.563867	1.27797	0.2538	-0.089695
10	0.760359	-1.551719	0.085434	-0.228784	-0.02557
11	-0.108804	-1.304662	0.015308	0.067615	0.190341
12	2.43732	1.782466	1.242653	0.360918	0.050018
13	1.099124	-0.021095	0.1285	0.257637	-0.013079
14	-1.065767	1.330131	0.360992	-0.677247	0.092075

	Pop	School	Employ	Health	Home
Pop	1	0.610194	0.970733	0.739984	-0.171965
School	0.610194	1	0.494304	0.095393	0.185928
Employ	0.970733	0.494304	1	0.847965	-0.249162
Health	0.739984	0.095393	0.847965	1	-0.357996
Home	-0.171965	0.185928	-0.249162	-0.357996	1

Variable dropped	Alpha	Change
none	0.744261	
Pop	0.495563	-0.248698
School	0.705055	-0.039206
Employ	0.607932	-0.136329
Health	0.679789	-0.064472
Home	0.837606	0.093345

Polynomial regression

Data are from McClave and Deitrich (1991) and are provided in the "Home Size" and "KW Hrs/Mnth" columns of the "test" workbook.

Polynomial regression

Intercept	b0= -1216.143887	t = -5.008698	P = 0.0016
Home Size	b1= 2.39893	t = 9.75827	P < 0.0001
Home Size^2	b2= -0.00045	t = -7.617907	P = 0.0001

KW Hrs/Mnth = -1216.143887 +2.39893 Home Size -0.00045 Home Size^2

Analysis of variance from regression

Source of variation	Sum Squares	DF	Mean Square
Regression	831069.546371	2	415534.773185
Residual	15332.553629	7	2190.364804
Total (corrected)	846402.1	9	

Root MSE = 46.801333

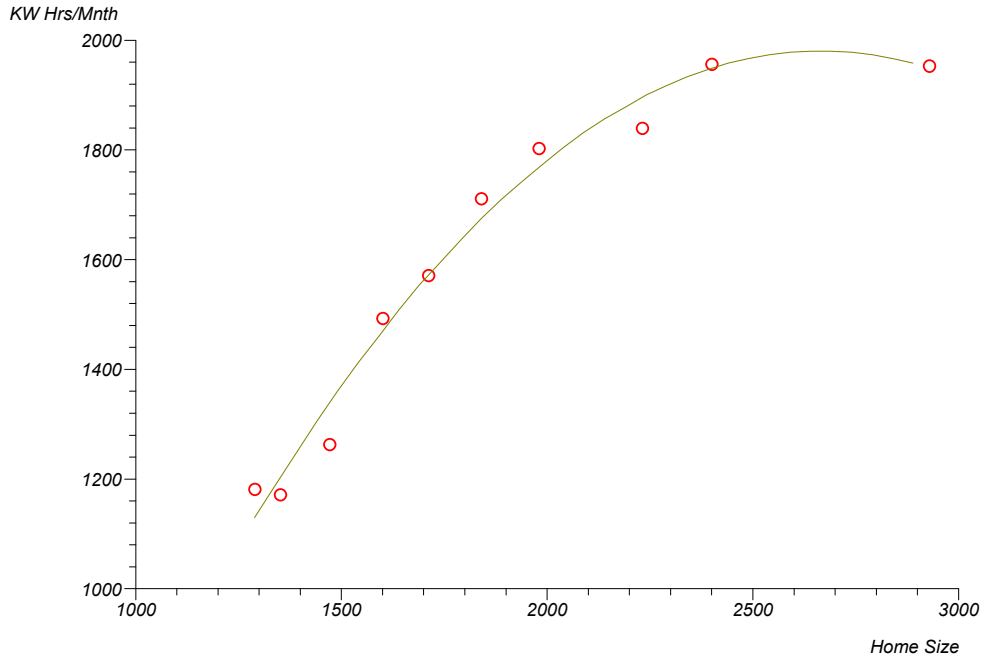
F = 189.710304 P < 0.0001

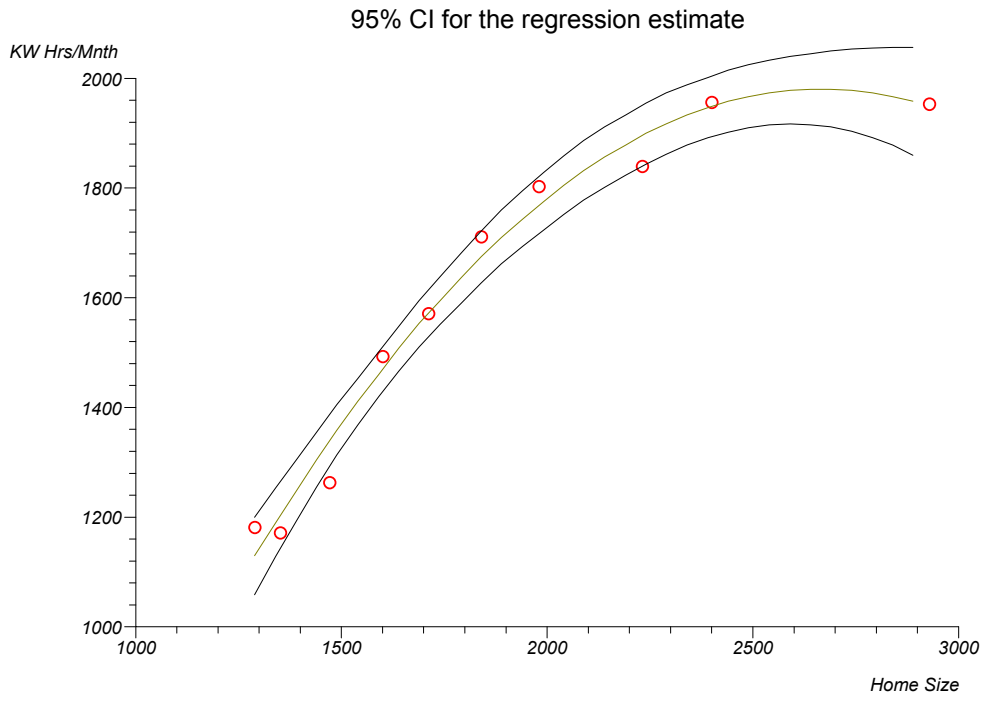
Multiple correlation coefficient (R) = 0.990901

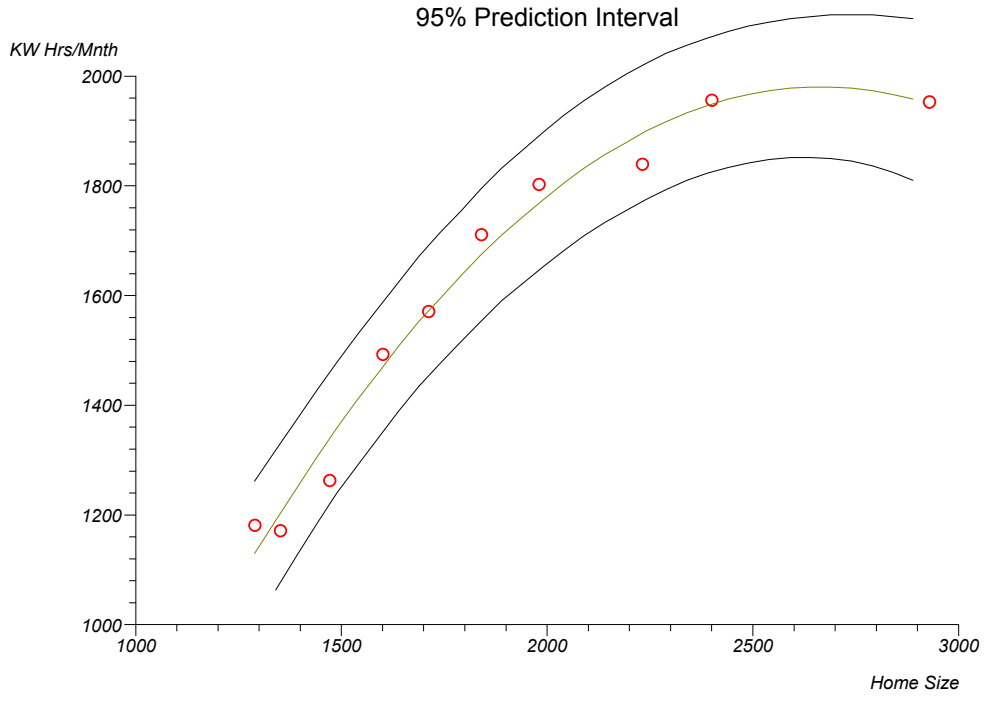
R^2 = 98.188502%

Ra^2 = 97.670932%

Durbin-Watson test statistic = 1.63341







Polynomial regression - area under curve

AUC (polynomial function) = 2855413.374801

AUC (by trapezoidal rule) = 2838195

Logistic regression

Data are from Altman (1991) and are provided in the "Men", "Hypertensive", "Smoking", "Obesity" and "Snoring" columns of the "test" workbook.

Logistic regression

Deviance (likelihood ratio) chi-square = 12.507498 df = 3 P = 0.0058

Intercept	b0 = -2.377661	z = -6.253967	P < 0.0001
Smoking	b1 = -0.067775	z = -0.243686	P = 0.8075
Obesity	b2 = 0.69531	z = 2.438954	P = 0.0147
Snoring	b3 = 0.871939	z = 2.193152	P = 0.0283

logit Hypertensive = -2.377661 -0.067775 Smoking +0.69531 Obesity +0.871939 Snoring

Logistic regression - model analysis

Accuracy = 1.00E-07
Log likelihood with all covariates = -199.4582

Deviance with all covariates = 1.618403 df = 4 rank = 4
Akaike = 9.618403
Schwartz = 12.01561

Deviance with no covariates = 14.1259

Deviance (likelihood ratio) chi-square = 12.507498 df = 3 P = 0.0058

Pearson chi-square goodness of fit = 1.364272 df = 4 P = 0.8504
Deviance goodness of fit = 1.618403 df = 4 P = 0.8055
Hosmer-Lemeshow test = 0.453725 df = 2 P = 0.797

Parameter	Coefficient	Standard Error
Constant	-2.377661	0.380185
Smoking	-0.067775	0.278124
Obesity	0.69531	0.285085
Snoring	0.871939	0.397574

Logistic regression - odds ratios

Parameter	Estimate	Odds Ratio	95% CI
Constant	-2.377661		
Smoking	-0.067775	0.934471	0.541784 to 1.611779
Obesity	0.69531	2.00433	1.146316 to 3.504564
Snoring	0.871939	2.391544	1.097143 to 5.213072

Logistic regression - classification

		Observed	
		Event	No Event
Classified	Event	0	0
	No Event	79	354

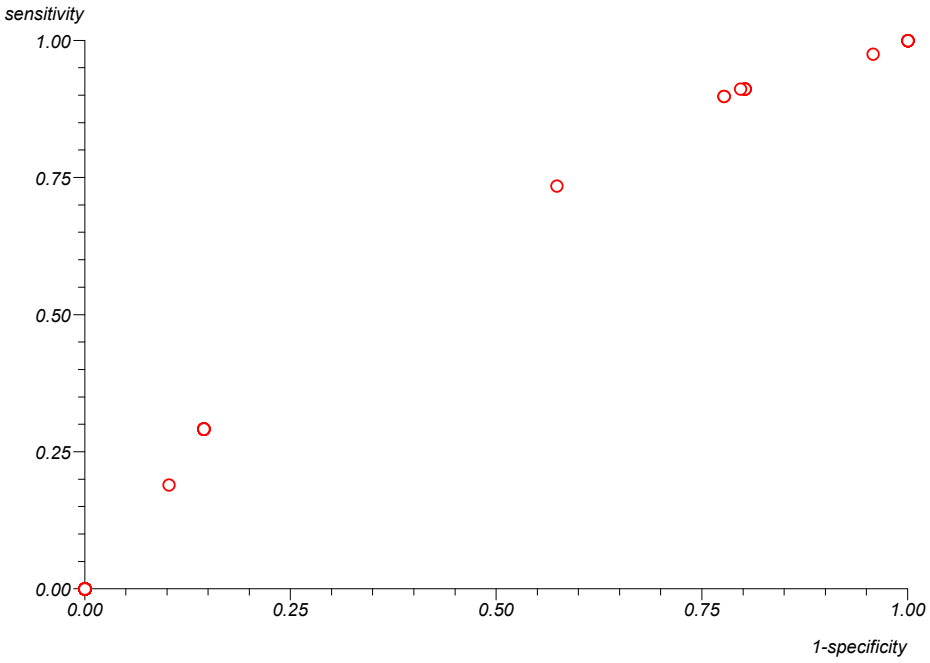
cutoff = 0.5

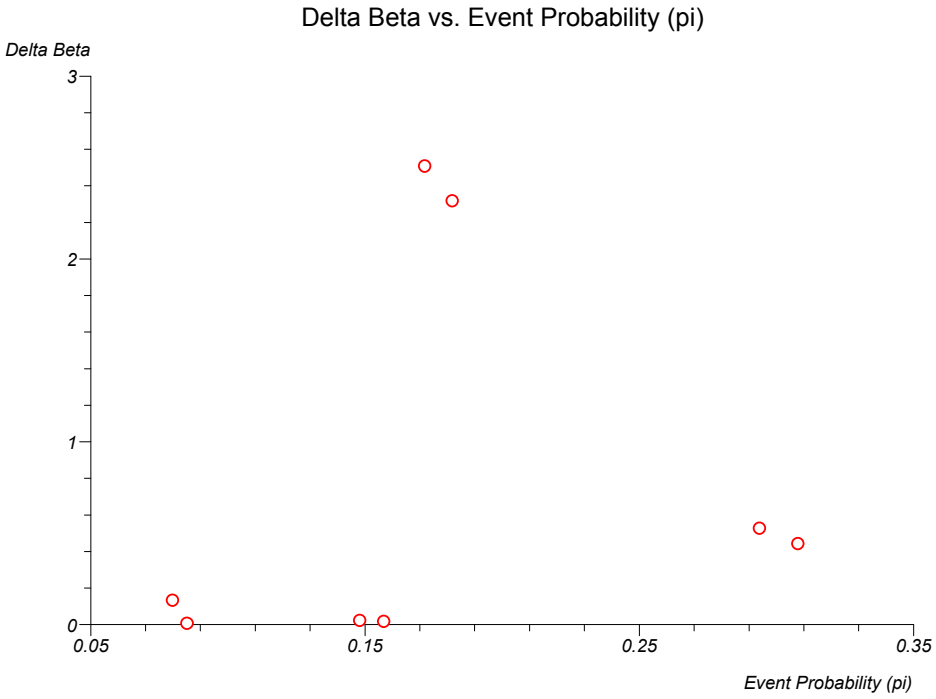
sensitivity = 0%
specificity = 100%

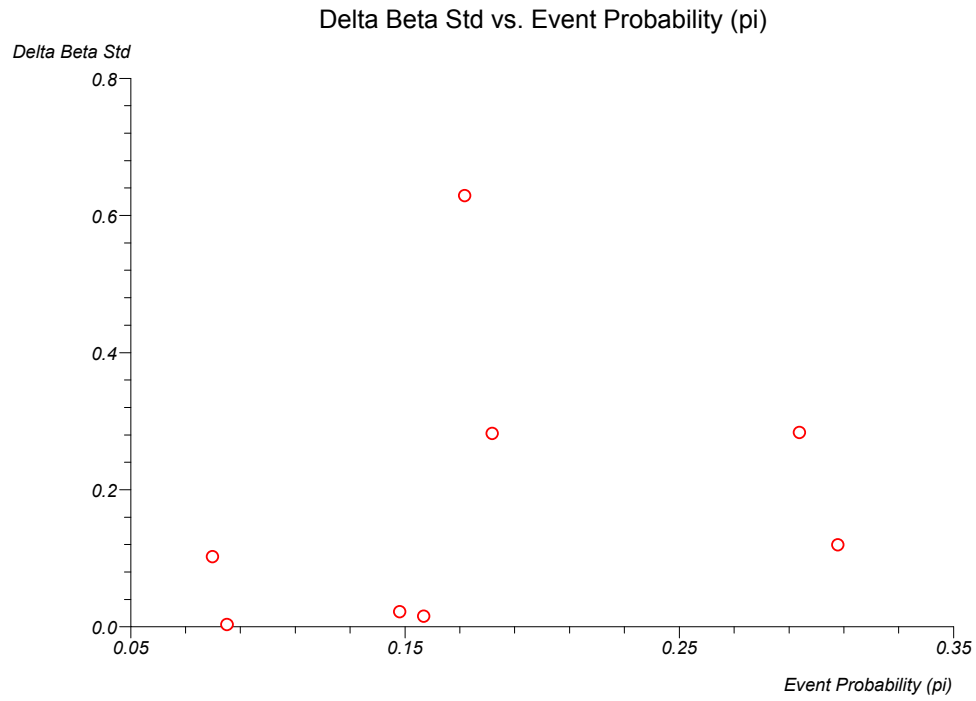
+ve predictive value = *%
-ve predictive value = 81.76%
predictive value despite -ve test = 18.24%

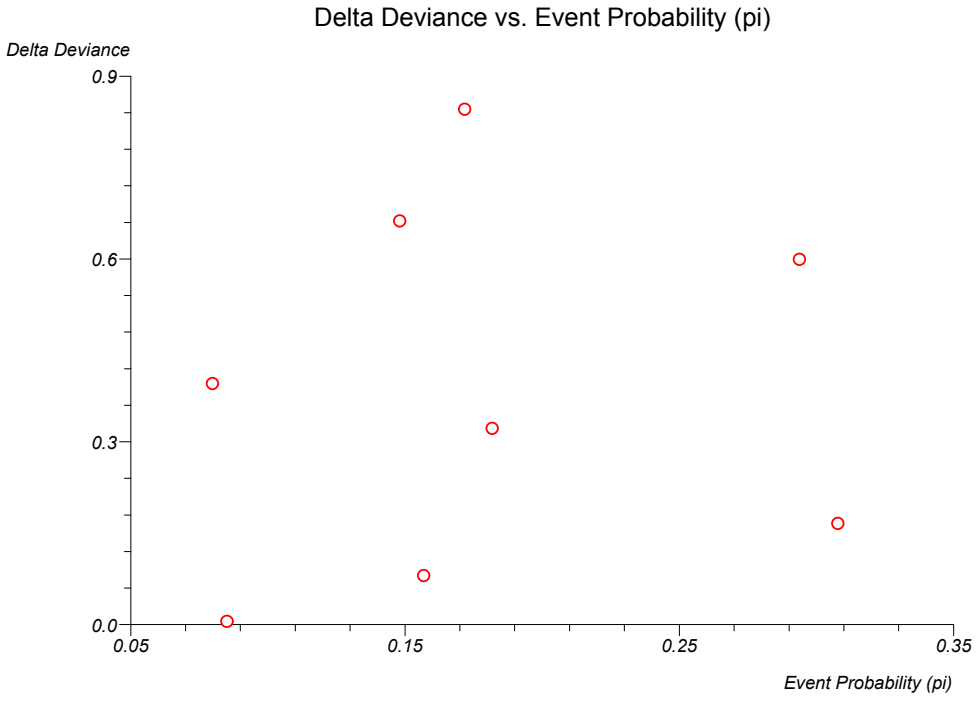
Likelihood ratio (95% CI):
LR (+ve) = * (* to *)
LR (-ve) = 1 (0.943864 to 1.007922)

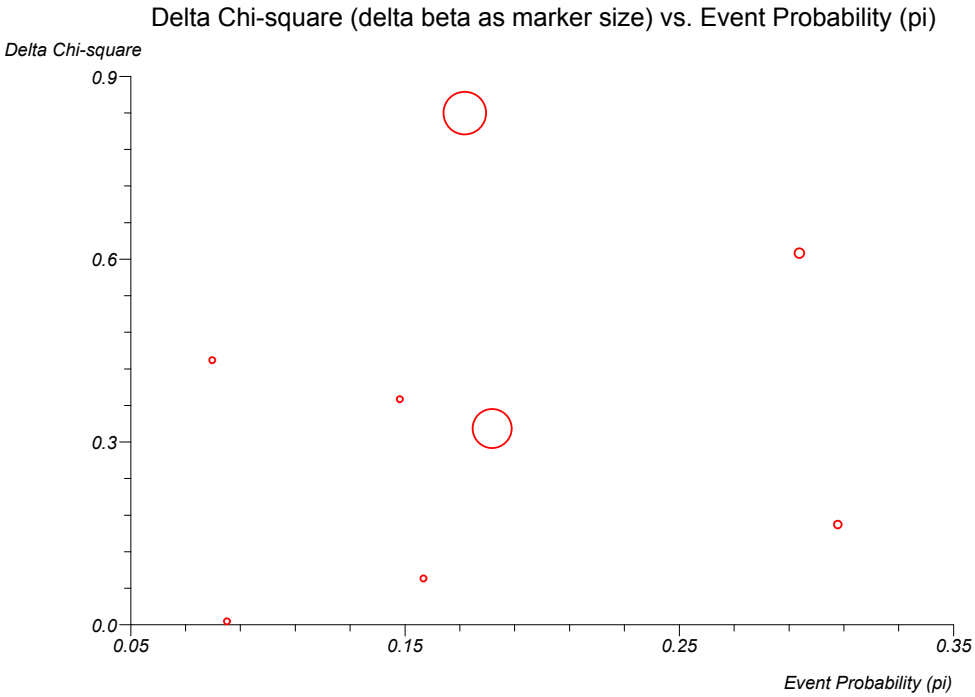
'Near' cut-off at max(sens+spec) = 0.18

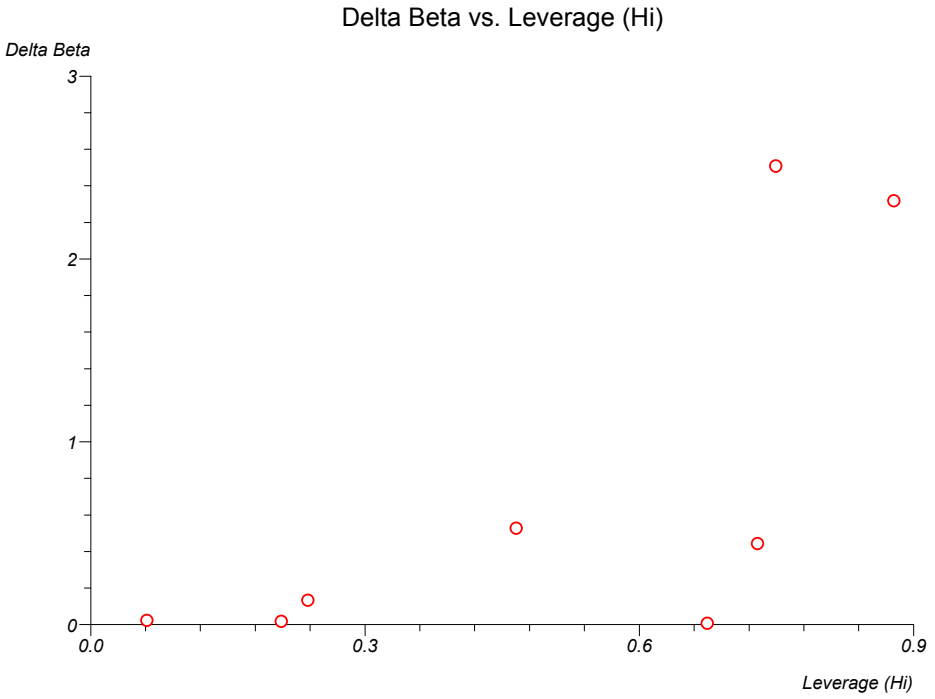

sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1-specificity

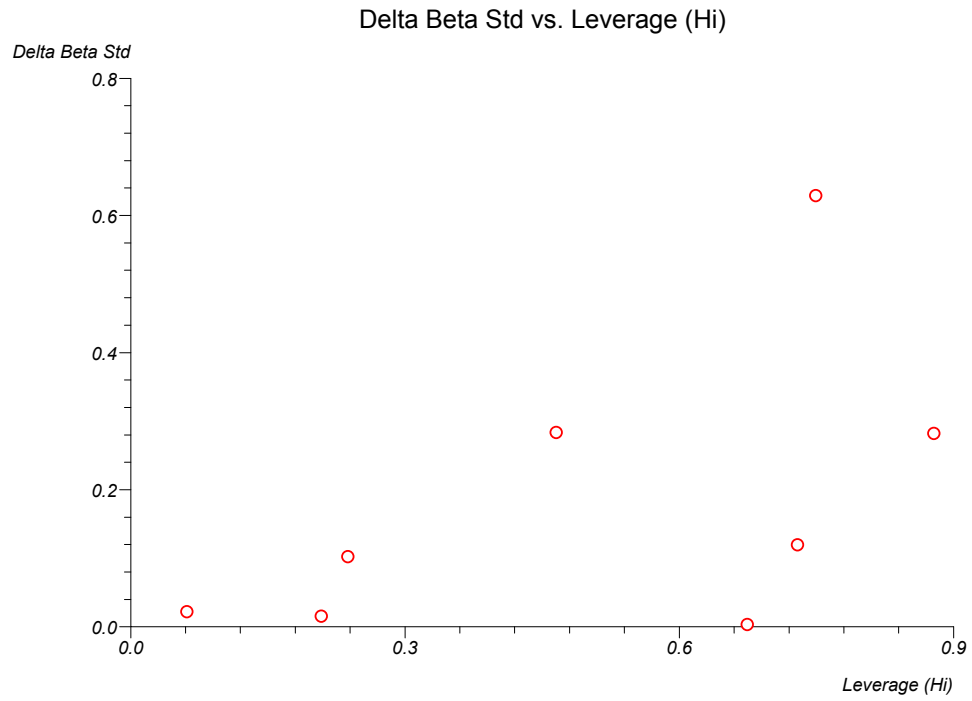

Delta Beta vs. Event Probability (pi)
Delta Beta
3
2
1
0
0.05
0.15
0.25
0.35
Event Probability (pi)

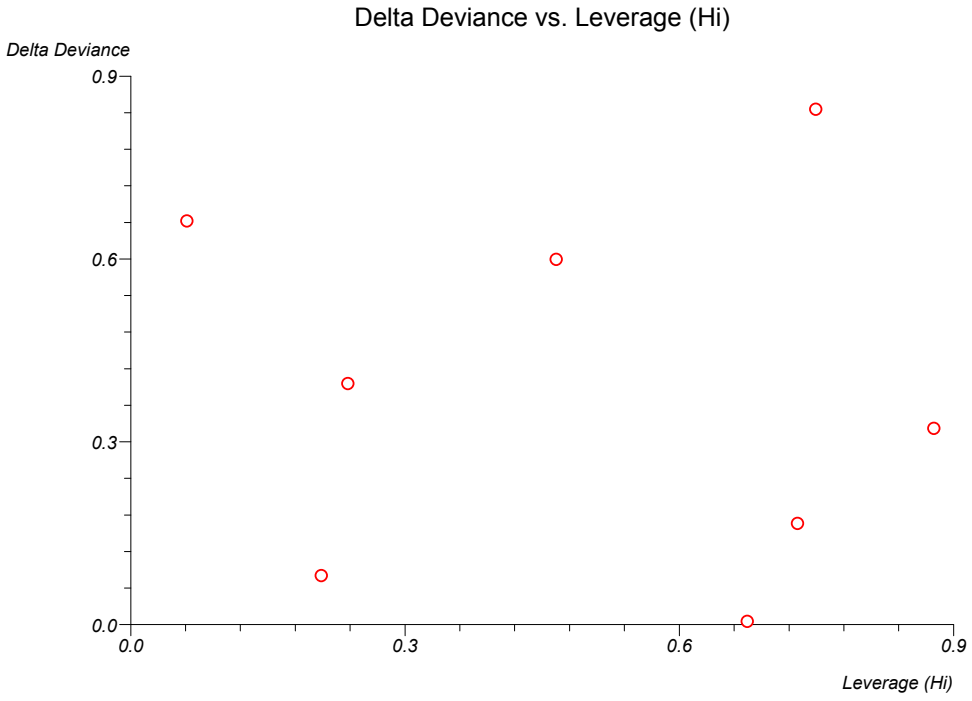

Delta Beta Std vs. Event Probability (pi)
Delta Beta Std
0.8
0.6
0.4
0.2
0.0
0.05
0.15
0.25
0.35
Event Probability (pi)

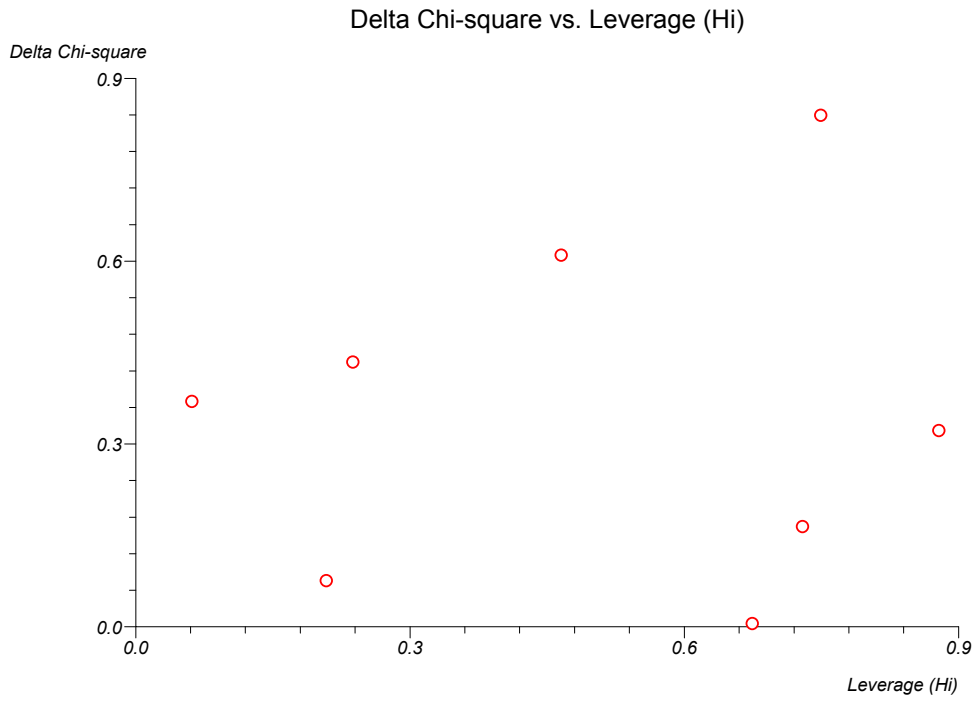

Delta Deviance vs. Event Probability (pi)
Delta Deviance
0.9
0.6
0.3
0.0
0.05
0.15
0.25
0.35
Event Probability (pi)


Delta Chi-square (delta beta as marker size) vs. Event Probability (pi)
Delta Chi-square
0.9
0.6
0.3
0.0
0.05
0.15
0.25
0.35
Event Probability (pi)


Delta Beta vs. Leverage (Hi)
Delta Beta
3
2
1
0
0.0
0.3
0.6
0.9
Leverage (Hi)


Delta Beta Std vs. Leverage (Hi)
Delta Beta Std
0.8
0.6
0.4
0.2
0.0
0.0
0.3
0.6
0.9
Leverage (Hi)


Delta Deviance vs. Leverage (Hi)
Delta Deviance
0.9
0.6
0.3
0.0
0.0
0.3
0.6
0.9
Leverage (Hi)



Logistic regression - fits and residuals

Index	Trials	Events	Event Probability	Deviance Residual
1	60	5	0.084892	-0.04344
2	17	2	0.079773	0.541452
3	8	1	0.156784	-0.254756
4	2	0	0.148031	-0.800513
5	187	35	0.181574	0.197592
6	85	13	0.171718	-0.466021
7	51	15	0.307803	-0.212624
8	23	8	0.293554	0.562314

Index	Pearson Residual	Leverage	Std Pearson Residual
1	-0.043319	0.67372	-0.075837
2	0.576358	0.236795	0.659738
3	-0.24725	0.207636	-0.277763
4	-0.589495	0.060918	-0.608314
5	0.198373	0.877953	0.56783
6	-0.459033	0.749169	-0.916545
7	-0.211727	0.729005	-0.406721
8	0.571559	0.464803	0.781275

Index	Delta Beta	Std Delta Beta	Delta Deviance	Delta Chi-square
1	0.003875	0.011875	0.005762	0.005751
2	0.103066	0.135044	0.396236	0.435254
3	0.01602	0.020218	0.08092	0.077152
4	0.022543	0.024005	0.663364	0.370046
5	0.283079	2.319415	0.322121	0.322431
6	0.629344	2.509038	0.846519	0.840055
7	0.120593	0.445003	0.165802	0.165422
8	0.283712	0.530107	0.599908	0.610391

Parameters	Covariance
Intercept vs. Intercept	0.14454
Intercept vs. Smoking	-0.016074
Intercept vs. Obesity	-0.014745
Intercept vs. Snoring	-0.135506
Smoking vs. Smoking	0.077353
Smoking vs. Obesity	-0.000008
Smoking vs. Snoring	-0.007416
Obesity vs. Obesity	0.081274
Obesity vs. Snoring	-0.008143
Snoring vs. Snoring	0.158065

Probit analysis

Data are from Finney (1971) and are provided in the "Age", "Girls" and "Menses" columns of the "test" workbook.

Probit analysis - logit sigmoid curve

constant = -10.613197

slope = 0.815984

Median * Dose = 13.006622

Confidence interval (No Heterogeneity) = 12.930535 to 13.082483

* Dose for centile 90 = 14.352986

Confidence interval (No Heterogeneity) = 14.238636 to 14.480677

Chi² (heterogeneity of deviations from model) = 21.869852 (23 df) P = 0.5281

t for slope = 27.682452 (23 df) P < 0.0001

Probit analysis - further statistics

Iterations = 4

Sxx = 1150.921353

Sxy = 939.13361

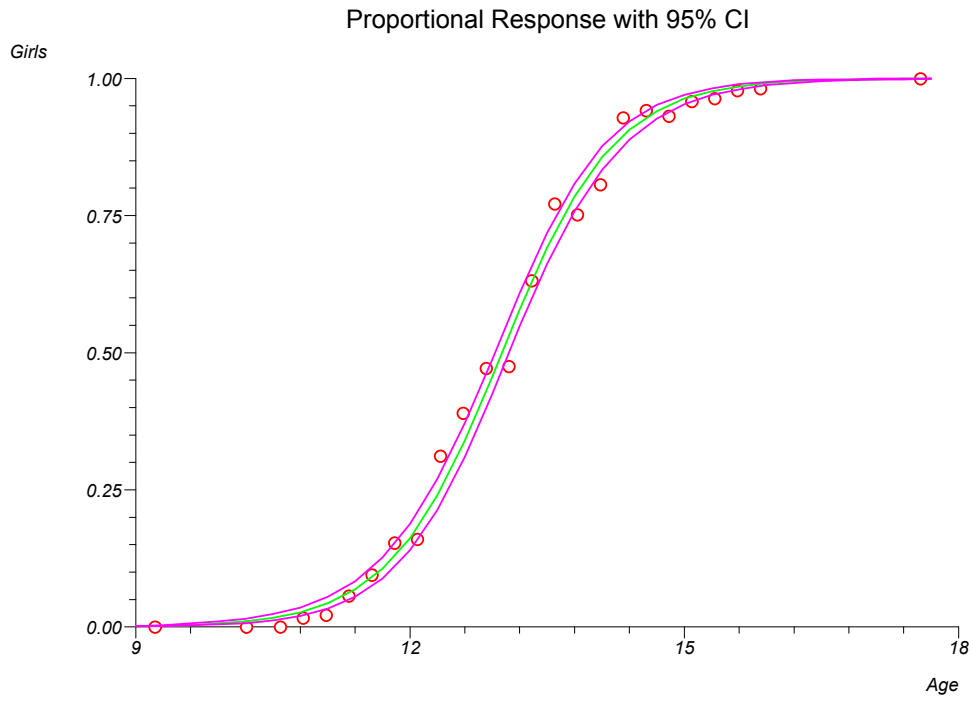
Syy = 788.188015

Variance of B = 0.000826

Standard error of B without heterogeneity = 0.029477

Index	Subjects	Responses	Expected	Deviation
1	376	0	0.764592	-0.764592
2	200	0	2.06257	-2.06257
3	93	0	1.739416	-1.739416
4	120	2	3.343623	-1.343623
5	90	2	3.718889	-1.718889
6	88	5	5.35666	-0.35666
7	105	10	9.325481	0.674519

8	111	17	14.190043	2.809957
9	100	16	18.061043	-2.061043
10	93	29	23.152263	5.847737
11	100	39	33.264793	5.735207
12	108	51	46.270932	4.729068
13	99	47	52.460303	-5.460303
14	106	67	66.669399	0.330601
15	105	81	75.414927	5.585073
16	117	88	92.792961	-4.792961
17	98	79	83.512615	-4.512615
18	97	90	86.967562	3.032438
19	120	113	111.450467	1.549533
20	102	95	97.049326	-2.049326
21	122	117	117.997281	-0.997281
22	111	107	108.551334	-1.551334
23	94	92	92.610792	-0.610792
24	114	112	112.874071	-0.874071
25	1049	1049	1048.398657	0.601343



Cox regression

Data are from Armitage and Berry (1994) and are provided in the "Time", "Censor" and "Stage group" columns of the "test" workbook.

Cox (proportional hazards) regression

Deviance (likelihood ratio) chi-square = 7.634383 df = 1 P = 0.0057

Stage group b1 = 0.96102 z = 2.492043 P = 0.0127

Cox regression - model analysis

Deviance with no covariates = 415.109602

Deviance with all model covariates = 407.475219

Deviance (likelihood ratio) chi-square = 7.634383 df = 1 P = 0.0057

Cox regression - coefficient detail

Parameter	Coefficient	Standard Error
Stage group	0.961020228844836	0.385635552349939

Parameter	95% CI	Risk Ratio (95% CI)
Stage group	0.205188 to 1.716852	2.614362 1.227756 to 5.566976

Cox regression - survival and hazard estimates (mean covariate)

Time	Survival probability	Cumulative hazard	Proportionality
4	0.988323	0.011745	1.25639
6	0.964972	0.035657	0.480572
10	0.953296	0.04783	1.25639
11	0.918046	0.085508	1.25639
13	0.906219	0.098475	1.25639
17	0.894352	0.111656	1.25639
19	0.882506	0.124989	0.480572
20	0.858815	0.152202	1.25639
21	0.846906	0.166166	1.25639

22	0.834954	0.180379	1.25639
24	0.810918	0.209589	1.25639
29	0.798831	0.224606	1.25639
30	0.774515	0.255519	1.25639
31	0.762283	0.271437	1.25639
32	0.750078	0.287578	0.480572
33	0.737899	0.303949	1.25639
34	0.725667	0.320664	1.25639
35	0.713383	0.337737	1.25639
39	0.701044	0.355185	1.25639
40	0.688649	0.373024	1.25639
42	0.66347	0.410272	0.480572
45	0.650549	0.429939	1.25639
46	0.637562	0.450104	1.25639
50	0.624506	0.470794	1.25639
56	0.611381	0.492035	1.25639
63	0.597425	0.515126	1.25639
68	0.583386	0.538906	1.25639
82	0.569259	0.563419	1.25639
85	0.555042	0.588711	1.25639
88	0.540731	0.614833	1.25639
89	0.526322	0.641841	1.25639
90	0.511812	0.669799	1.25639
93	0.497195	0.698774	1.25639
94	0.482635	0.728494	0.480572
104	0.468132	0.759005	1.25639
110	0.453513	0.790732	1.25639
134	0.438538	0.82431	1.25639
137	0.42343	0.859367	1.25639
169	0.407693	0.89724	1.25639
171	0.391239	0.938436	1.25639
173	0.374606	0.981881	1.25639
175	0.357781	1.027835	1.25639
184	0.34075	1.076605	1.25639
201	0.323499	1.12856	1.25639
207	0.306373	1.182954	0.480572
222	0.288899	1.241677	1.25639
253	0.26778	1.317588	0.480572

Cox regression - survival and hazard estimates (baseline)

Time	Survival probability	Cumulative hazard	Sum[exp(bx)]
4	0.985352	0.014757	1
6	0.95619	0.044799	0.382502
10	0.941677	0.060093	1
11	0.898138	0.107432	1
13	0.883625	0.123723	1
17	0.869112	0.140283	1
19	0.854674	0.157036	0.382502
20	0.825946	0.191225	1
21	0.811583	0.208769	1
22	0.797219	0.226626	1
24	0.768492	0.263325	1
29	0.754128	0.282193	1
30	0.725401	0.321031	1
31	0.711037	0.341031	1
32	0.696763	0.36131	0.382502
33	0.682578	0.381878	1
34	0.668393	0.402879	1
35	0.654208	0.424329	1
39	0.640023	0.446251	1
40	0.625838	0.468663	1
42	0.597225	0.515461	0.382502
45	0.582649	0.540171	1
46	0.568072	0.565506	1
50	0.553496	0.5915	1
56	0.53892	0.618188	1
63	0.52351	0.647199	1
68	0.5081	0.677076	1
82	0.492691	0.707874	1
85	0.477281	0.73965	1
88	0.461871	0.77247	1
89	0.446461	0.806403	1
90	0.431051	0.841528	1
93	0.415642	0.877932	1
94	0.400408	0.915272	0.382502
104	0.385349	0.953606	1
110	0.37029	0.993467	1

134	0.354994	1.035654	1
137	0.339697	1.0797	1
169	0.323912	1.127283	1
171	0.307574	1.179041	1
173	0.291235	1.233625	1
175	0.274896	1.291361	1
184	0.258558	1.352636	1
201	0.242219	1.417912	1
207	0.226219	1.486251	0.382502
222	0.21013	1.56003	1
253	0.191015	1.655404	0.382502

Cox regression - residuals and diagnostics

Index	Leverage	Proportionality	Cox-Oakes residual
1	0.118317	0.480572	0.016976
2	0.116485	0.480572	0.059317
3	0.112211	0.480572	0.136399
4	0.109496	0.480572	0.194765
5	0.109496	0.480572	0.194765
6	*	0.480572	0.194765
7	0.098936	0.480572	0.345487
8	*	0.480572	0.37482
9	*	0.480572	0.425139
10	0.081875	0.480572	0.558916
11	*	0.480572	0.558916
12	*	0.480572	0.586121
13	0.083482	0.480572	0.621945
14	*	0.480572	0.621945
15	*	0.480572	0.621945
16	*	0.480572	0.621945
17	*	0.480572	0.621945
18	*	0.480572	0.621945
19	*	0.480572	0.621945
20	0.001685	1.25639	0.014648
21	0.001736	1.25639	0.04438
22	0.001624	1.25639	0.059558
23	0.001674	1.25639	0.105794
24	0.001674	1.25639	0.105794

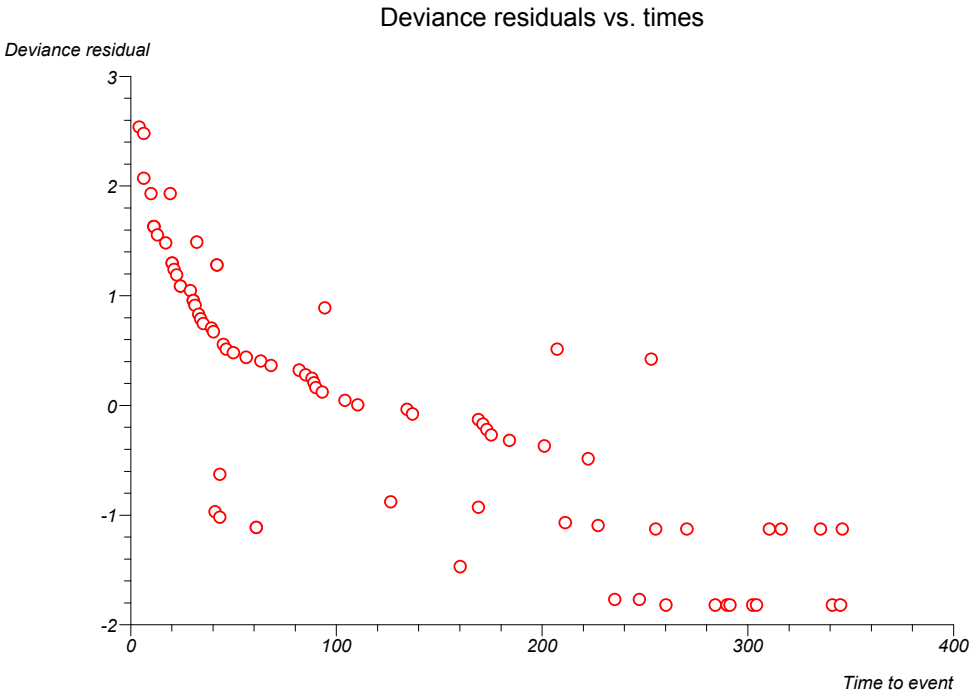
25	0.001674	1.25639	0.105794
26	0.001841	1.25639	0.121953
27	0.001902	1.25639	0.138377
28	0.001776	1.25639	0.188688
29	0.001776	1.25639	0.188688
30	0.001902	1.25639	0.206078
31	0.00197	1.25639	0.223777
32	0.002041	1.25639	0.259811
33	0.002041	1.25639	0.259811
34	0.002197	1.25639	0.278502
35	0.002281	1.25639	0.316595
36	0.002281	1.25639	0.316595
37	0.002465	1.25639	0.336396
38	0.002309	1.25639	0.376955
39	0.002406	1.25639	0.397737
40	0.002509	1.25639	0.418959
41	0.002619	1.25639	0.440642
42	0.002736	1.25639	0.462805
43	*	1.25639	0.462805
44	*	1.25639	0.509187
45	0.00219	1.25639	0.533594
46	0.002301	1.25639	0.558611
47	0.002421	1.25639	0.58427
48	0.00255	1.25639	0.610605
49	*	1.25639	0.610605
50	*	1.25639	0.610605
51	0.003006	1.25639	0.639199
52	0.003186	1.25639	0.668634
53	0.003382	1.25639	0.698962
54	0.003597	1.25639	0.730239
55	0.003833	1.25639	0.762526
56	0.004093	1.25639	0.79589
57	0.004381	1.25639	0.830405
58	0.004699	1.25639	0.866155
59	0.004431	1.25639	0.940837
60	0.004785	1.25639	0.979915
61	0.004493	1.25639	1.021225
62	0.004888	1.25639	1.064314
63	*	1.25639	1.064314

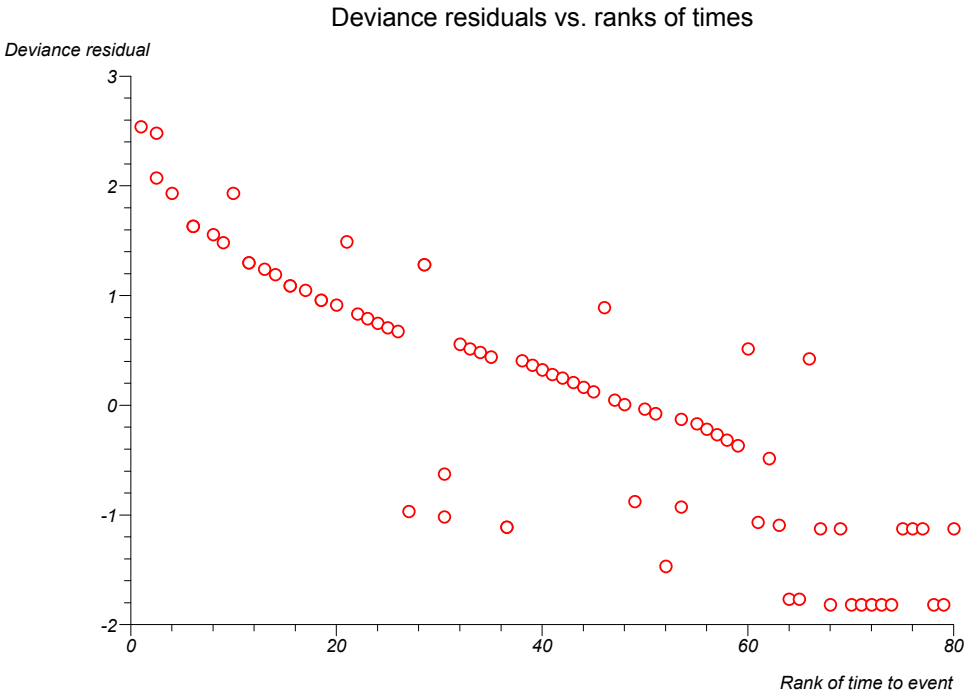
64	0.005854	1.25639	1.111467
65	0.005536	1.25639	1.161908
66	0.00614	1.25639	1.215029
67	0.006848	1.25639	1.27113
68	0.007686	1.25639	1.330565
69	0.008688	1.25639	1.393757
70	0.007044	1.25639	1.532334
71	*	1.25639	1.532334
72	*	1.25639	1.532334
73	*	1.25639	1.625988
74	*	1.25639	1.625988
75	*	1.25639	1.625988
76	*	1.25639	1.625988
77	*	1.25639	1.625988
78	*	1.25639	1.625988
79	*	1.25639	1.625988
80	*	1.25639	1.625988

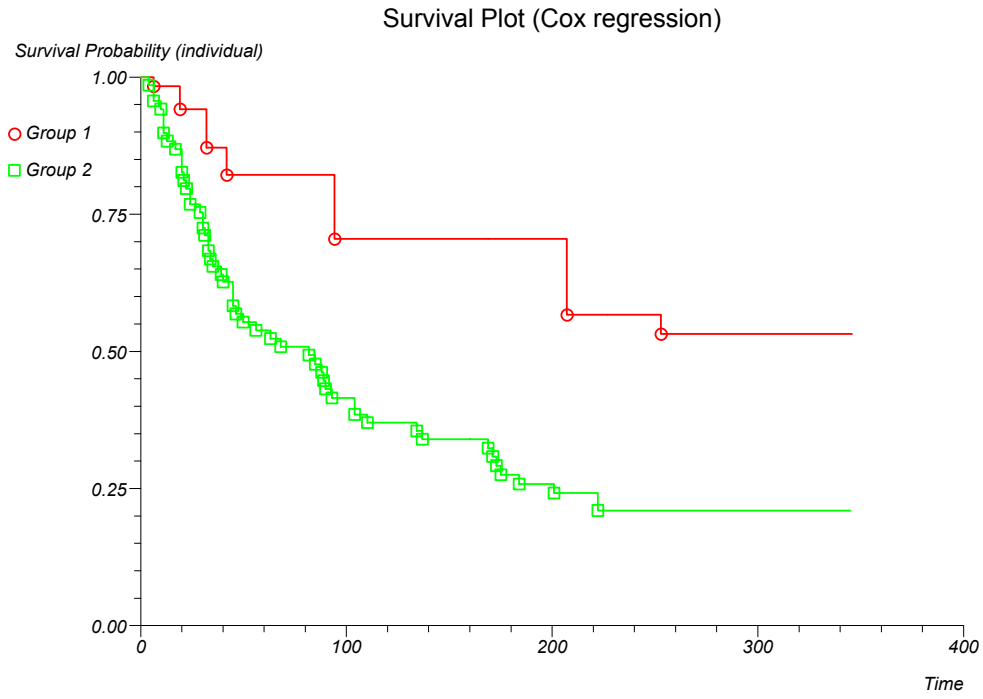
Index	Cox-Snell residual	Martingale residual	Deviance residual
1	0.014757	0.985243	2.541976
2	0.017136	0.982864	2.483439
3	0.044799	0.955201	2.073826
4	0.060093	0.939907	1.934919
5	0.107432	0.892568	1.636052
6	0.107432	0.892568	1.636052
7	0.107432	0.892568	1.636052
8	0.123723	0.876277	1.557842
9	0.140283	0.859717	1.486186
10	0.060066	0.939934	1.935133
11	0.191225	0.808775	1.300406
12	0.191225	0.808775	1.300406
13	0.208769	0.791231	1.245228
14	0.226626	0.773374	1.192544
15	0.263325	0.736675	1.093335
16	0.263325	0.736675	1.093335
17	0.282193	0.717807	1.046286
18	0.321031	0.678969	0.956294
19	0.321031	0.678969	0.956294
20	0.341031	0.658969	0.913032

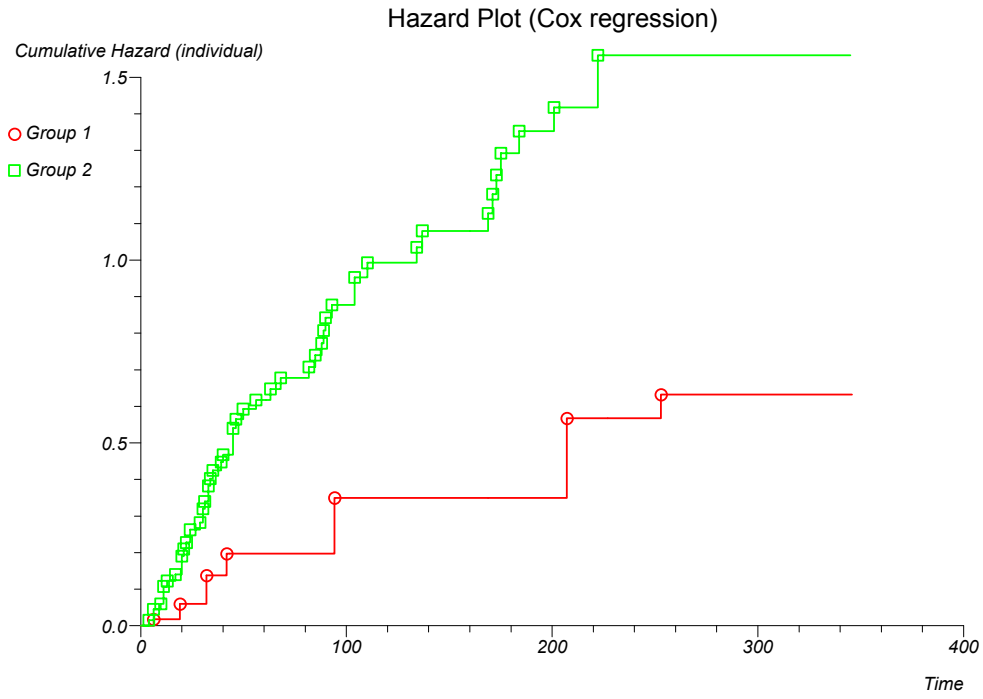
21	0.138202	0.861798	1.494819
22	0.381878	0.618122	0.830098
23	0.402879	0.597121	0.789935
24	0.424329	0.575671	0.750433
25	0.446251	0.553749	0.711513
26	0.468663	0.531337	0.673104
27	0.468663	-0.468663	-0.968156
28	0.197165	0.802835	1.28131
29	0.197165	0.802835	1.28131
30	0.197165	-0.197165	-0.627957
31	0.515461	-0.515461	-1.015344
32	0.540171	0.459829	0.558642
33	0.565506	0.434494	0.520654
34	0.5915	0.4085	0.482894
35	0.618188	0.381812	0.44531
36	0.618188	-0.618188	-1.111925
37	0.618188	-0.618188	-1.111925
38	0.647199	0.352801	0.405711
39	0.677076	0.322924	0.36619
40	0.707874	0.292126	0.32669
41	0.73965	0.26035	0.287152
42	0.77247	0.22753	0.247516
43	0.806403	0.193597	0.207724
44	0.841528	0.158472	0.167714
45	0.877932	0.122068	0.127422
46	0.350094	0.649906	0.894033
47	0.953606	0.046394	0.047132
48	0.993467	0.006533	0.006547
49	0.380004	-0.380004	-0.871784
50	1.035654	-0.035654	-0.035239
51	1.0797	-0.0797	-0.077676
52	1.0797	-1.0797	-1.46949
53	0.431189	-0.431189	-0.928643
54	1.127283	-0.127283	-0.122252
55	1.179041	-0.179041	-0.16935
56	1.233625	-0.233625	-0.217568
57	1.291361	-0.291361	-0.267074
58	1.352636	-0.352636	-0.318059
59	1.417912	-0.417912	-0.370747

60	0.568495	0.431505	0.516252
61	0.568495	-0.568495	-1.066297
62	1.56003	-0.56003	-0.48026
63	0.596715	-0.596715	-1.092443
64	1.56003	-1.56003	-1.766369
65	1.56003	-1.56003	-1.766369
66	0.633196	0.366804	0.424668
67	0.633196	-0.633196	-1.125341
68	1.655404	-1.655404	-1.819562
69	0.633196	-0.633196	-1.125341
70	1.655404	-1.655404	-1.819562
71	1.655404	-1.655404	-1.819562
72	1.655404	-1.655404	-1.819562
73	1.655404	-1.655404	-1.819562
74	1.655404	-1.655404	-1.819562
75	0.633196	-0.633196	-1.125341
76	0.633196	-0.633196	-1.125341
77	0.633196	-0.633196	-1.125341
78	1.655404	-1.655404	-1.819562
79	1.655404	-1.655404	-1.819562
80	0.633196	-0.633196	-1.125341


Deviance residuals vs. times
Deviance residual
3
2
1
0
-1
-2
0
100
200
300
400
Time to event


Deviance residuals vs. ranks of times
Deviance residual
3
2
1
0
-1
-2
0
20
40
60
80
Rank of time to event


Survival Plot (Cox regression)
Survival Probability (individual)
1.00
0.75
0.50
0.25
0.00
Group 1
Group 2
0
100
200
300
400
Time


Hazard Plot (Cox regression)
Cumulative Hazard (individual)
1.5
1.0
0.5
0.0
Group 1
Group 2
0
100
200
300
400
Time

Kaplan-Meier survival estimates

Data are from Kalbfleisch and Prentice (1980) and are provided in the "Group Surv", "Time Surv" and "Censor Surv" columns of the "test" workbook.

Kaplan-Meier survival estimates

Group: 1 (Group Surv = 2)

Time	At risk	Dead	Censored	S	SE(S)	H	SE(H)
142	22	1	0	0.954545	0.044409	0.04652	0.046524
157	21	1	0	0.909091	0.061291	0.09531	0.06742
163	20	1	0	0.863636	0.073165	0.146603	0.084717
198	19	1	0	0.818182	0.08223	0.200671	0.100504
204	18	0	1	0.818182	0.08223	0.200671	0.100504
205	17	1	0	0.770053	0.090387	0.261295	0.117378
232	16	3	0	0.625668	0.105069	0.468935	0.16793
233	13	4	0	0.433155	0.108192	0.836659	0.249777
239	9	1	0	0.385027	0.106338	0.954442	0.276184
240	8	1	0	0.336898	0.103365	1.087974	0.306814
261	7	1	0	0.28877	0.099172	1.242125	0.34343
280	6	2	0	0.192513	0.086369	1.64759	0.44864
295	4	2	0	0.096257	0.064663	2.340737	0.671772
323	2	1	0	0.048128	0.046941	3.033884	0.975335
344	1	0	1	0.048128	0.046941	3.033884	0.975335

Median survival time = 233

- Andersen 95% CI for median survival time = 231.898503 to 234.101497
- Brookmeyer-Crowley 95% CI for median survival time = 232 to 240

Mean survival time (95% CI) [limit: 344 on 323] = 241.283422 (219.591463 to 262.975382)

Group: 2 (Group Surv = 1)

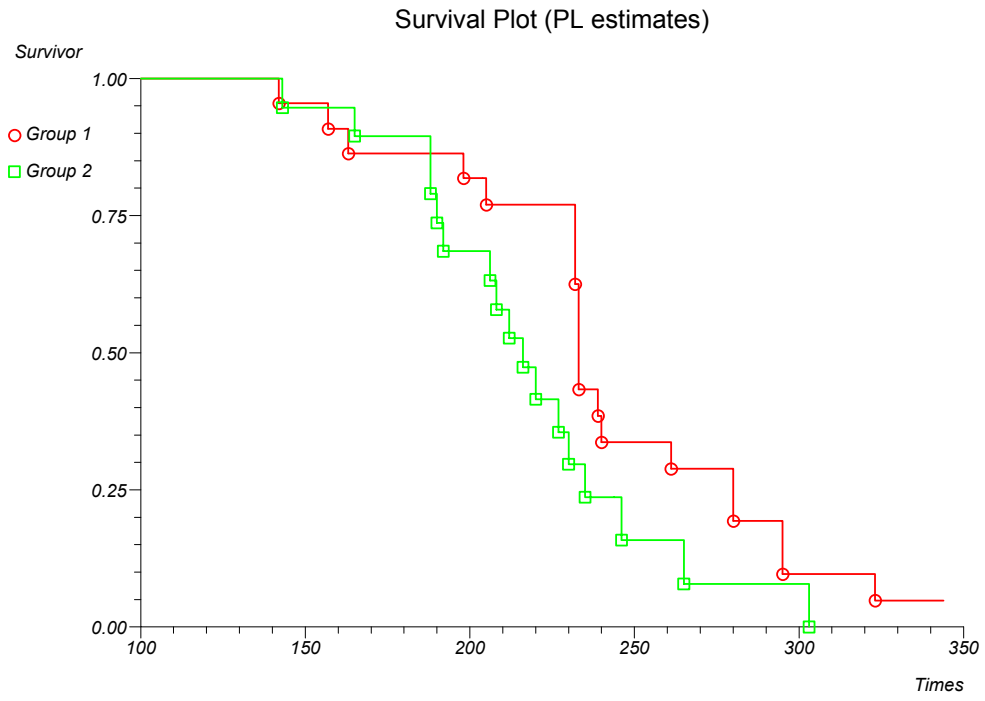
Time	At risk	Dead	Censored	S	SE(S)	H	SE(H)
143	19	1	0	0.947368	0.051228	0.054067	0.054074
165	18	1	0	0.894737	0.070406	0.111226	0.078689
188	17	2	0	0.789474	0.093529	0.236389	0.11847
190	15	1	0	0.736842	0.101023	0.305382	0.137102
192	14	1	0	0.684211	0.106639	0.37949	0.155857

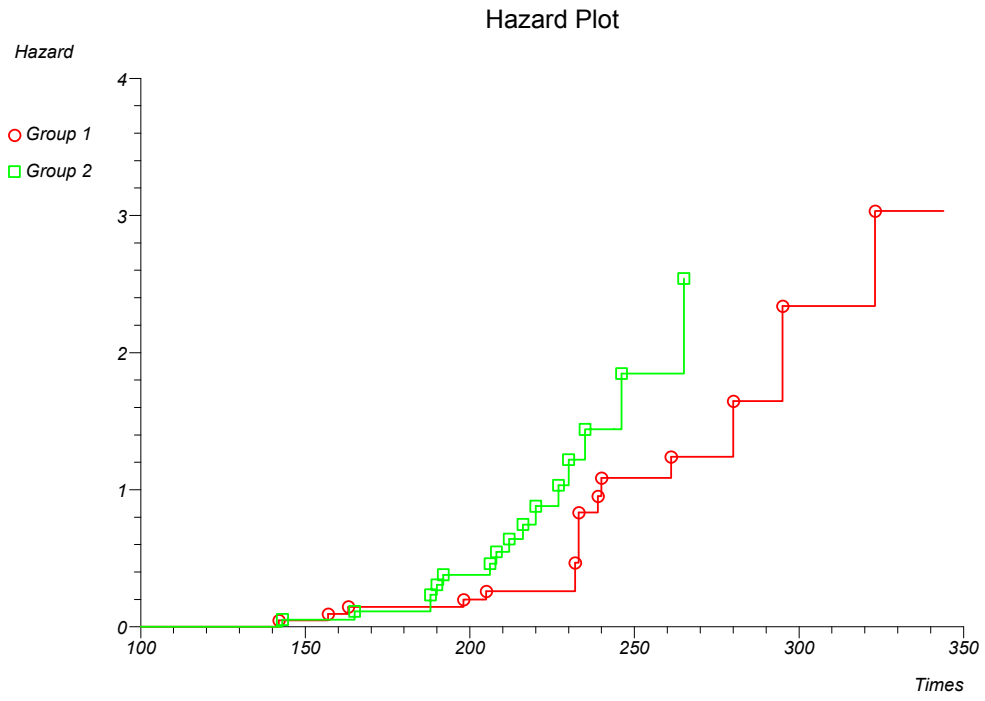
206	13	1	0	0.631579	0.110665	0.459532	0.175219
208	12	1	0	0.578947	0.113269	0.546544	0.195646
212	11	1	0	0.526316	0.114549	0.641854	0.217643
216	10	1	1	0.473684	0.114549	0.747214	0.241825
220	8	1	0	0.414474	0.114515	0.880746	0.276291
227	7	1	0	0.355263	0.112426	1.034896	0.316459
230	6	1	0	0.296053	0.108162	1.217218	0.365349
235	5	1	0	0.236842	0.10145	1.440362	0.428345
244	4	0	1	0.236842	0.10145	1.440362	0.428345
246	3	1	0	0.157895	0.093431	1.845827	0.591732
265	2	1	0	0.078947	0.072792	2.538974	0.922034
303	1	1	0	0	*	infinity	*

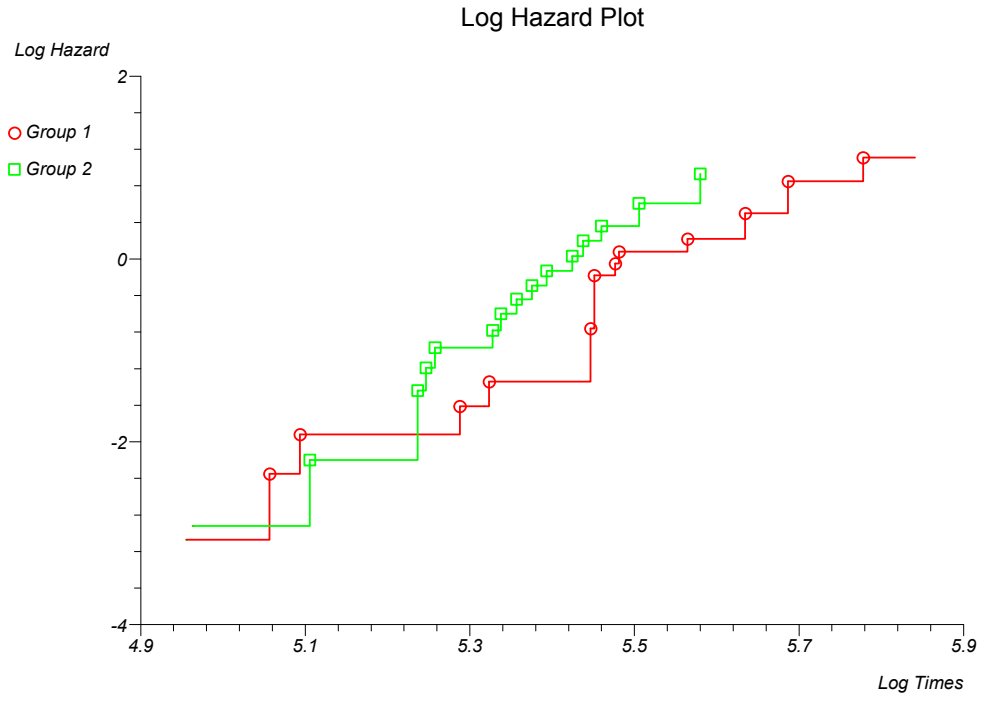
Median survival time = 216

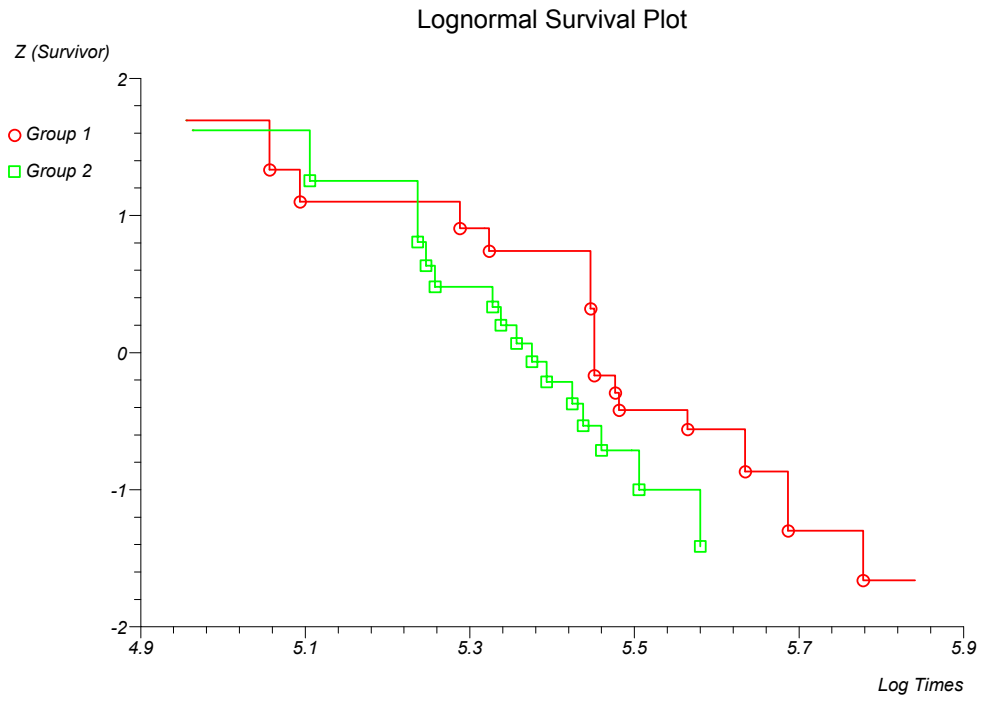
- Andersen 95% CI for median survival time = 199.619628 to 232.380372
- Brookmeyer-Crowley 95% CI for median survival time = 192 to 230

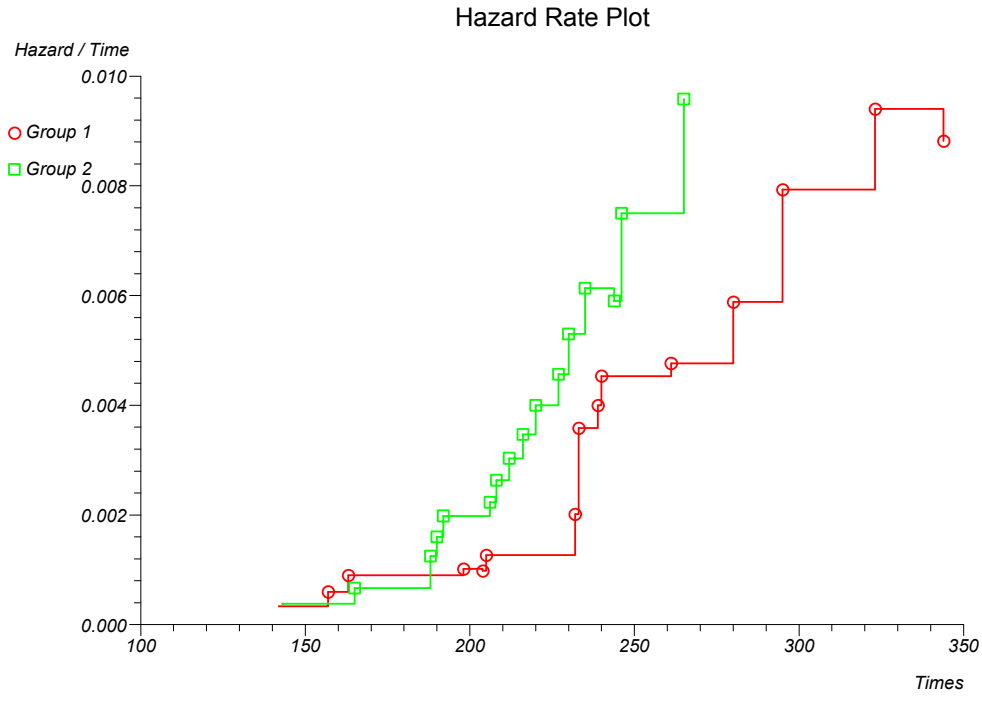
Mean survival time (95% CI) = 218.684211 (200.363485 to 237.004936)




Hazard Plot
Hazard
Group 1
Group 2
4
3
2
1
0
100
150
200
250
300
350
Times


Log Hazard Plot
Log Hazard
Group 1
Group 2
2
0
-2
-4
4.9
5.1
5.3
5.5
5.7
5.9
Log Times


Lognormal Survival Plot
Z (Survivor)
Group 1
Group 2
2
1
0
-1
-2
4.9
5.1
5.3
5.5
5.7
5.9
Log Times


Hazard Rate Plot
Hazard / Time
Group 1
Group 2
0.010
0.008
0.006
0.004
0.002
0.000
100
150
200
250
300
350
Times

Life table

Data are from Armitage and Berry (1994) and provided in the "Year", "Died" and "Withdrawn" columns.

Simple life table

Interval	Deaths	Withdrawn	At risk	Adj. at risk	P(death)
0 to 1	90	0	374	374	0.240642
1 to 2	76	0	284	284	0.267606
2 to 3	51	0	208	208	0.245192
3 to 4	25	12	157	151	0.165563
4 to 5	20	5	120	117.5	0.170213
5 to 6	7	9	95	90.5	0.077348
6 to 7	4	9	79	74.5	0.053691
7 to 8	1	3	66	64.5	0.015504
8 to 9	3	5	62	59.5	0.05042
9 to 10	2	5	54	51.5	0.038835
10 up	21	26	47	*	*

Interval	P(survival)	Survivors (lx%)	SD of lx%	95% CI for lx%
0 to 1	0.759358	100	*	* to *
1 to 2	0.732394	75.935829	10.57424	71.271289 to 79.951252
2 to 3	0.754808	55.614973	7.87331	50.428392 to 60.482341
3 to 4	0.834437	41.97861	7.003571	36.945565 to 46.922332
4 to 5	0.829787	35.028509	6.747202	30.200182 to 39.889161
5 to 6	0.922652	29.066209	6.651959	24.47156 to 33.805
6 to 7	0.946309	26.817994	6.659494	22.322081 to 31.504059
7 to 8	0.984496	25.378102	6.700832	20.935141 to 30.043836
8 to 9	0.94958	24.984643	6.720449	20.552912 to 29.648834
9 to 10	0.961165	23.724913	6.803396	19.323326 to 28.39237
10 up	*	22.803557	6.886886	18.417247 to 27.483099

Log-rank and Wilcoxon comparisons of survival

The two sample unstratified data are from Armitage and Berry (1994) and are provided in the "Stage group", "Time" and "Censor" columns of the "test" workbook. The two sample stratified data are from Peto et al. (1977) and are provided in the "Group", "Trial Time", "Censorship" and "Strat" columns of the "test" workbook. The four sample data are from Hosmer and Lemeshow (1999) and are provided in the "Age Groups (HMO)", "Times (HMO)" and "Censor (HMO)" columns of the "test" workbook.

Unstratified two sample example:

Log-rank and Wilcoxon tests

Log-rank (Peto):

For group 1 (Group Surv = 2)
Observed deaths = 20
Extent of exposure to risk of death = 24.796316
Relative rate = 0.806571

For group 2 (Group Surv = 1)
Observed deaths = 17
Extent of exposure to risk of death = 12.203684
Relative rate = 1.393022

test statistics:
-4.796316 4.796316

variance-covariance matrix:
0.135829 -7.362217
-7.362217 7.362217

Chi-square for equivalence of death rates = 3.12469 P = 0.0771

Hazard Ratio (approximate 95% confidence interval)
Group 1 vs. Group 2 = 0.579008 (0.291772 to 1.149018)

Generalised Wilcoxon (Peto-Prentice):

test statistics:
-3.005122 3.005122

variance-covariance matrix:
0.323989 -3.08652
-3.08652 3.08652

Chi-square for equivalence of death rates = 2.925871 P = 0.0872

Stratified two sample example:

Log-rank and Wilcoxon tests

Log-rank (Peto) * [STRATUM 1 of 2]:

For group 1 (Group = 1)
Observed deaths = 4
Extent of exposure to risk of death = 5.421429
Relative rate = 0.737813

For group 2 (Group = 2)
Observed deaths = 3
Extent of exposure to risk of death = 1.578571
Relative rate = 1.900452

test statistics:
-1.421429 1.421429

variance-covariance matrix:
1.084131 -0.922398
-0.922398 0.922398

Chi-square for equivalence of death rates = 2.190442 P = 0.1389

Generalised Wilcoxon (Peto-Prentice) * [STRATUM 1 of 2]:

test statistics:
-0.75 0.75

variance-covariance matrix:
2.133333 -0.46875
-0.46875 0.46875

Chi-square for equivalence of death rates = 1.2 P = 0.2733

Log-rank (Peto) * [STRATUM 2 of 2]:

For group 1 (Group = 1)
Observed deaths = 2
Extent of exposure to risk of death = 4.98342
Relative rate = 0.401331

For group 2 (Group = 2)
Observed deaths = 8
Extent of exposure to risk of death = 5.01658
Relative rate = 1.594712

test statistics:
-2.983422.98342

variance-covariance matrix:
0.410957 -2.433343
-2.433343 2.433343

Chi-square for equivalence of death rates = 3.657846 P = 0.0558

Log-rank (Peto) for COMBINED STRATA:

Stratum	Deaths	Extent of exposure to risk of death	Relative rate
1	6	10.404848	0.576654
2	11	6.595152	1.667892

overall chi-square = 5.781939 P = 0.0162

Hazard Ratio (approximate 95% confidence interval)

Group 1 vs. Group 2 = 0.345738 (0.13034 to 0.9171)

Generalised Wilcoxon (Peto-Prentice) * [STRATUM 2 of 2]:

test statistics:

-1.885167 1.885167

variance-covariance matrix:

0.776548 -1.287751

-1.287751 1.287751

Chi-square for equivalence of death rates = 2.759739 P = 0.0967

Generalised Wilcoxon (Peto-Prentice) for COMBINED STRATA:

Stratum	Deaths	Extent of exposure to risk of death	Relative rate
1	6	10.404848	0.576654
2	11	6.595152	1.667892

overall chi-square = 3.953375 P = 0.0468

Unstratified k sample example:

Using group scores of 25, 32.5, 37.5 and 47.5.

Log-rank and Wilcoxon tests

For group 1 (Age Group (HMO) = 4)
Observed deaths = 23
Extent of exposure to risk of death = 12.866192
Relative rate = 1.787631

For group 2 (Age Group (HMO) = 3)
Observed deaths = 20
Extent of exposure to risk of death = 17.81137
Relative rate = 1.122878

For group 3 (Age Group (HMO) = 2)
Observed deaths = 29
Extent of exposure to risk of death = 29.434373
Relative rate = 0.985243

For group 4 (Age Group (HMO) = 1)
Observed deaths = 8
Extent of exposure to risk of death = 19.888065
Relative rate = 0.402251

test statistics:

10.133808	2.18863	-0.434373	-11.888065

variance-covariance matrix:

0.164164	0.07133	0.071164	-2.054433
0.07133	0.129208	0.068062	-3.600897
0.071164	0.068062	0.106726	-5.703517
-2.054433	-3.600897	-5.703517	11.358847

Chi-square for equivalence of death rates = 19.905841 P = 0.0002
Chi-square for trend = 19.293476 P < 0.0001

Hazard Ratio (approximate 95% confidence interval)

Group 1 vs. Group 2 = 1.592008 (0.777157 to 3.26123)

Group 1 vs. Group 3 = 1.814406 (0.94244 to 3.493136)

Group 1 vs. Group 4 = 4.444064 (2.204144 to 8.960264)

Group 2 vs. Group 3 = 1.139697 (0.632793 to 2.05266)

Group 2 vs. Group 4 = 2.791484 (1.472816 to 5.290805)

Group 3 vs. Group 4 = 2.449321 (1.386658 to 4.326354)

Generalised Wilcoxon (Gehan-Breslow):

test statistics:

6.831683	0.762376	-3.019802	-4.574257

variance-covariance matrix:

0.4247	0.263118	0.260057	-0.855656
0.263118	0.408799	0.258988	-0.959892
0.260057	0.258988	0.358139	-1.476746
-0.855656	-0.959892	-1.476746	3.292293

Chi-square for equivalence of death rates = 14.143278 P = 0.0027

Chi-square for trend = 13.935225 P = 0.0002

Wei-Lachin test

Data are from Makuch and Escobar (1991) and are provided in the "Treatment Gp", "time m1", "censor m1", "time m2", "censor m2", "time m3", "censor m3", "time m4" and "censor m4" columns of the test workbook.

Wei-Lachin Analysis

Univariate Generalised Wilcoxon (Gehan)

total cases = 47 (by group = 23 and 24)

Repeat time 1

observed failures by group =	20 and 23	
Wei-Lachin t =	-0.527597	
Wei-Lachin variance =	0.077575	
chi-square =	3.588261	P = 0.0582

Repeat time 2

observed failures by group =	14 and 21	
Wei-Lachin t =	0.077588	
Wei-Lachin variance =	0.056161	
chi-square =	0.107189	P = 0.7434

Repeat time 3

observed failures by group =	18 and 19	
Wei-Lachin t =	-0.11483	
Wei-Lachin variance =	0.060918	
chi-square =	0.216452	P = 0.6418

Repeat time 4

observed failures by group =	20 and 16	
Wei-Lachin t =	0.335179	
Wei-Lachin variance =	0.056281	
chi-square =	1.996143	P = 0.1577

Multivariate Generalised Wilcoxon (Gehan)

Covariance matrix:

0.077575			
0.026009	0.056161		
0.035568	0.020484	0.060918	
0.023525	0.016862	0.026842	0.056281

Inverse of covariance matrix:

19.204259			
-5.078483	22.22316		
-8.40436	-3.176864	25.857118	
-2.497583	-3.020025	-7.867237	23.468861

repeat times = 4

chi-square omnibus statistic = 9.242916 P = 0.0553

stochastic ordering chi-square = 0.095982 P = 0.7567

Univariate Log-Rank

total cases = 47 (by group = 23 and 24)

Repeat time 1

observed failures by group =	20 and 23	
Wei-Lachin t =	-0.716191	
Wei-Lachin variance =	0.153385	
chi-square =	3.344058	P = 0.0674

Repeat time 2

observed failures by group =	14 and 21	
Wei-Lachin t =	-0.277786	
Wei-Lachin variance =	0.144359	
chi-square =	0.534536	P = 0.4647

Repeat time 3

observed failures by group =	18 and 19	
Wei-Lachin t =	-0.372015	
Wei-Lachin variance =	0.150764	
chi-square =	0.917956	P = 0.338

Repeat time 4

observed failures by group =	20 and 16	
Wei-Lachin t =	0.619506	
Wei-Lachin variance =	0.143437	
chi-square =	2.675657	P = 0.1019

Multivariate Log-Rank

Covariance matrix:

0.153385			
0.049439	0.144359		
0.052895	0.050305	0.150764	
0.039073	0.047118	0.052531	0.143437

Inverse of covariance matrix:

7.973385			
-1.779359	8.69056		
-1.892007	-1.661697	8.575636	
-0.894576	-1.761494	-2.079402	8.555558

repeat times = 4

chi-square omnibus statistic = 9.52966 P = 0.0491

stochastic ordering chi-square = 0.474382 P = 0.491

Odds ratio meta-analysis

Data are from Armitage and Berry (1994) and are provided in the "Smokers total", "Smokers cancer", "Control total" and "Control cancer" columns of the "test" workbook.

Odds ratio meta-analysis

Stratum	Table (a, b, c, d)			
1	83	3	72	14
2	90	3	227	43
3	129	7	81	19
4	412	32	299	131
5	1350	7	1296	61
6	60	3	106	27
7	459	18	534	81
8	499	19	462	56
9	451	39	1729	636
10	260	5	259	28

Stratum	Odds ratio	95% CI (Gart exact)		M-H Weight
1	5.37963	1.409677	30.08388	1.255814
2	5.682819	1.74247	29.261952	1.876033
3	4.322751	1.639419	12.650626	2.402542
4	5.640886	3.682557	8.815935	10.947368
5	9.077381	4.126151	23.59936	3.342668
6	5.09434	1.463436	27.155805	1.622449
7	3.867978	2.256644	6.950972	8.802198
8	3.183413	1.828153	5.757667	8.472973
9	4.253771	3.018787	6.136299	23.618564
10	5.621622	2.093208	18.88813	2.346014

Mantel-Haenszel chi-square = 292.379352 $P < 0.0001$

Mantel-Haenszel pooled estimate of odds ratio = 4.681639

Using the Robins, Breslow and Greenland method:

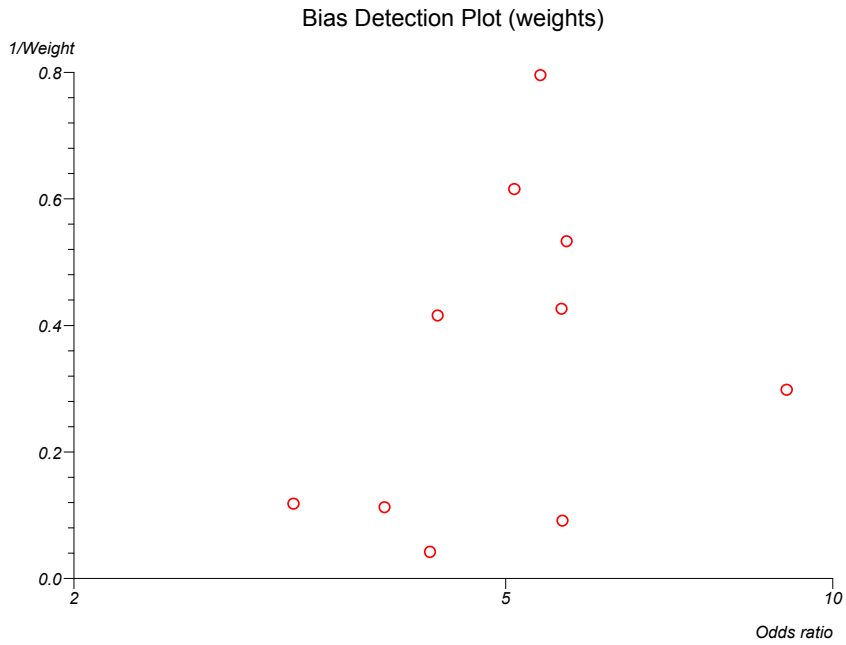
Approximate 95% CI = 3.865935 to 5.669455

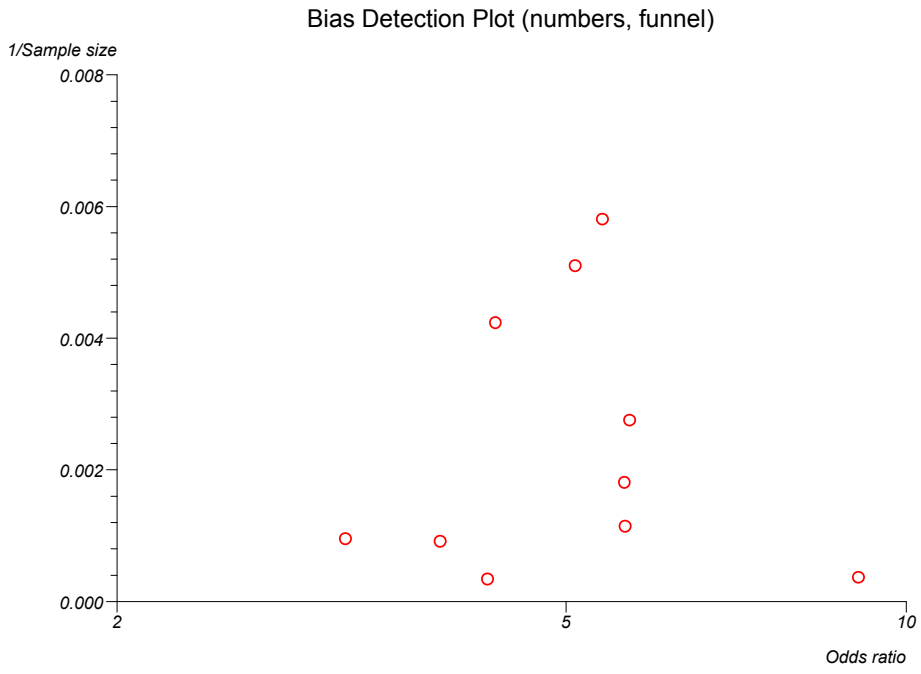
Q ("non-combinability" for odds ratios) = 6.641235 (df = 9) P = 0.6744

DerSimonian-Laird pooled odds ratio = 4.625084

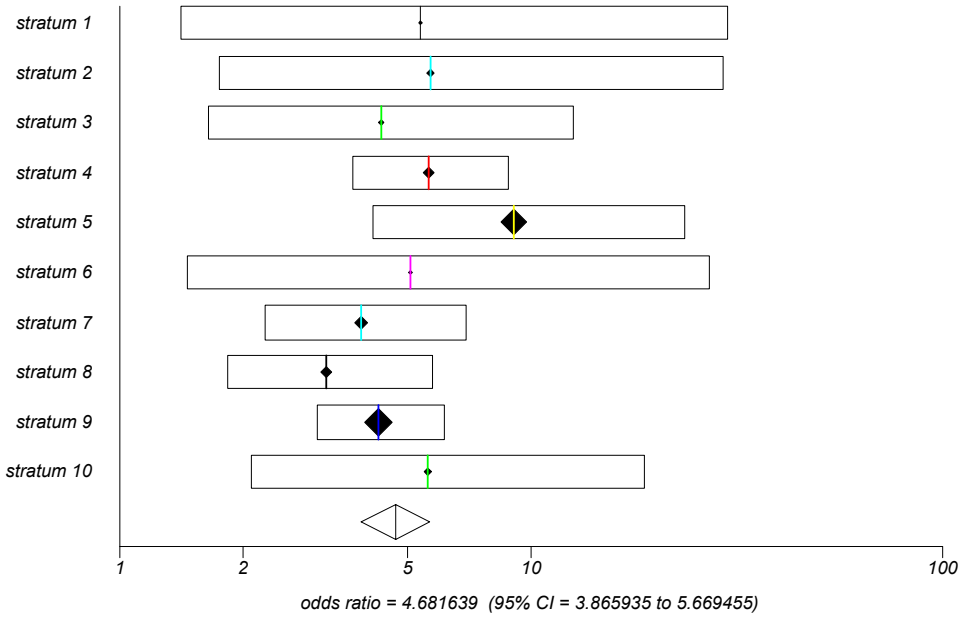
Approximate 95% CI = 3.821652 to 5.597423

DerSimonian-Laird chi-square = 247.466729 (df = 1) P < 0.0001

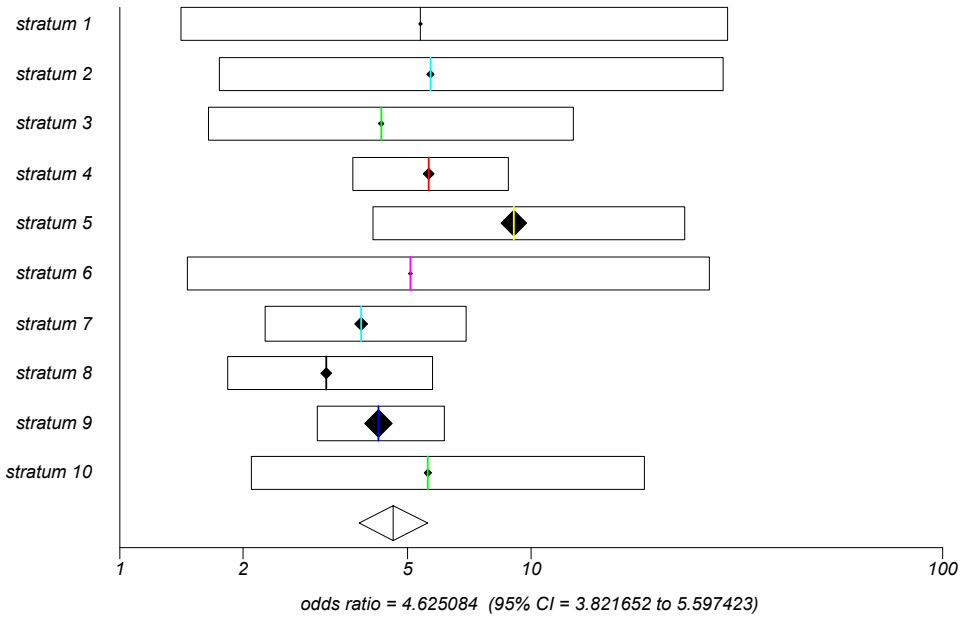




Cochrane odds ratio plot (fixed effects)



Cochrane odds ratio plot (random effects)



Peto odds ratio meta-analysis

Data are from Fleiss (1993) and Fleiss and Gross (1991), and are provided in the "Exposed total", "Exposed cases", "Non-exposed total", "Non-exposed cases" and "Study" columns of the "test" workbook.

Peto odds ratio meta-analysis

Stratum	Table (a, b, c, d)				
1	49	67	566	557	MRC-1
2	44	64	714	707	CDP
3	102	126	730	724	MRC-2
4	32	38	285	271	GASP
5	85	52	725	354	PARIS
6	246	219	2021	2038	AMIS
7	1570	1720	7017	6880	ISIS-2

Stratum	O-E	Odds ratio	95% CI		Peto weight (V)	
1	-8.578692	0.721713	0.492493	1.057619	26.304751	MRC-1
2	-9.540876	0.68386	0.462479	1.011214	25.107478	CDP
3	-10.780024	0.803583	0.607852	1.062341	49.29715	MRC-2
4	-3.447284	0.80134	0.487603	1.316943	15.565457	GASP
5	-6.258224	0.793516	0.544402	1.156621	27.058869	PARIS
6	12.986074	1.132558	0.934809	1.372138	104.323777	AMIS
7	-73.755746	0.895032	0.829531	0.965705	665.092282	ISIS-2

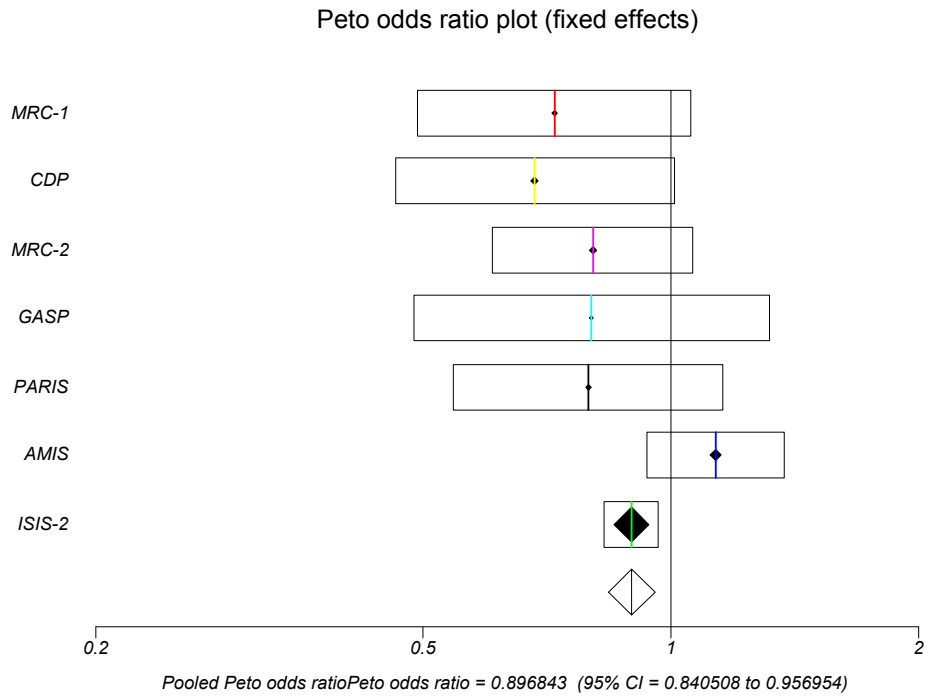
Stratum	z	P(two sided)	
1	-1.672646	P = 0.0944	MRC-1
2	-1.904087	P = 0.0569	CDP
3	-1.535355	P = 0.1247	MRC-2
4	-0.873768	P = 0.3822	GASP
5	-1.203085	P = 0.2289	PARIS
6	1.271412	P = 0.2036	AMIS
7	-2.859927	P = 0.0042	ISIS-2

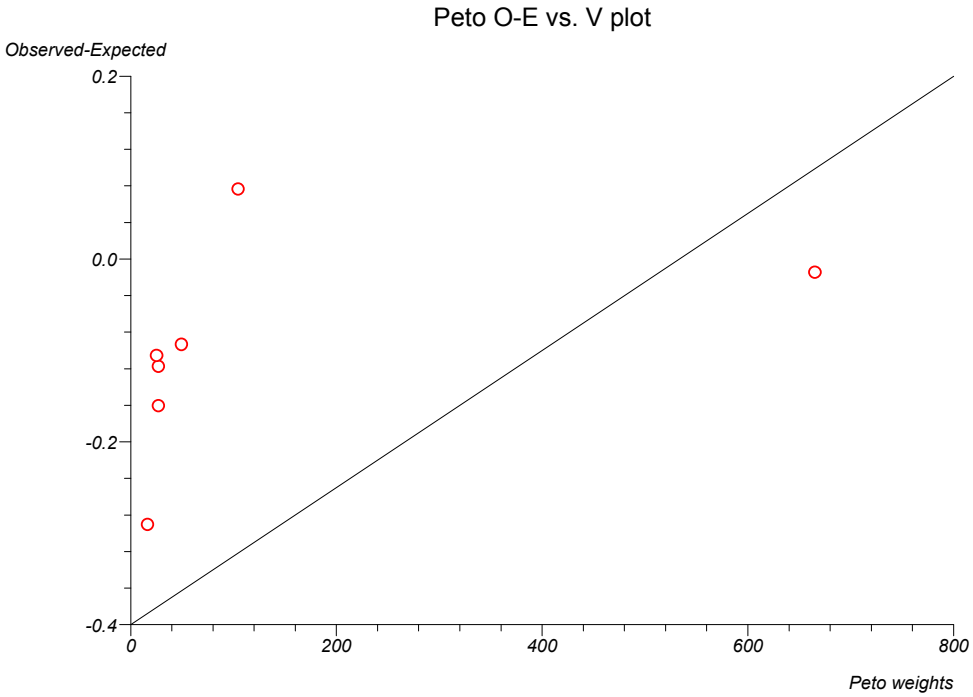
Pooled odds ratio = 0.896843

Approximate 95% CI = 0.840508 to 0.956954

Test statistic z for odds ratio = -3.289276 two sided P = 0.001

Test statistic for "non-combinability" = 9.967823 P = 0.126





Relative risk meta-analysis

Data are from Fleiss and Gross (1991) and are provided in the "Exposed total", "Exposed cases", "Non-exposed total", "Non-exposed cases" and "Study" columns of the "test" workbook.

Relative risk meta-analysis

Stratum	Table (a, b, c, d)				
1	49	67	566	557	MRC-1
2	44	64	714	707	CDP
3	102	126	730	724	MRC-2
4	32	38	285	271	GASP
5	85	52	725	354	PARIS
6	246	219	2021	2038	AMIS
7	1570	1720	7017	6880	ISIS-2

Stratum	Relative risk	95% CI (near exact)		M-H weight	
1	0.742046	0.522928	1.051866	33.256659	MRC-1
2	0.699291	0.483538	1.010302	31.727927	CDP
3	0.827038	0.648842	1.053557	62.325803	MRC-2
4	0.820853	0.528416	1.273886	19.242812	GASP
5	0.819326	0.594653	1.133252	34.638158	PARIS
6	1.118333	0.941378	1.3287	109.742042	AMIS
7	0.914173	0.859614	0.97217	859.349508	ISIS-2

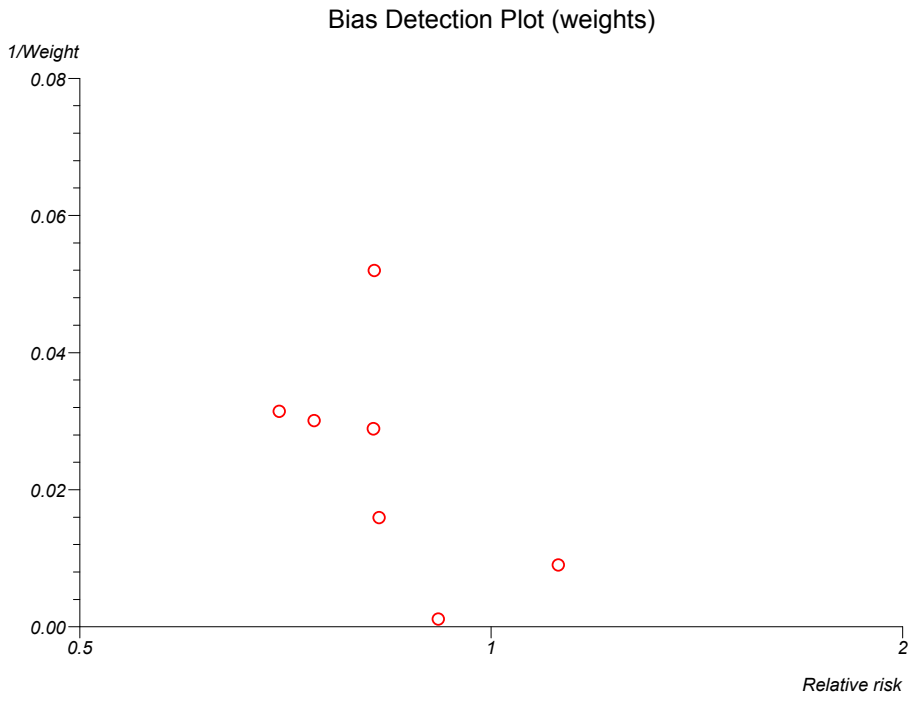
M-H pooled estimate (Rothman-Boice) of relative risk = 0.913608
Robins-Greenland approximate 95% CI = 0.8657 to 0.964168

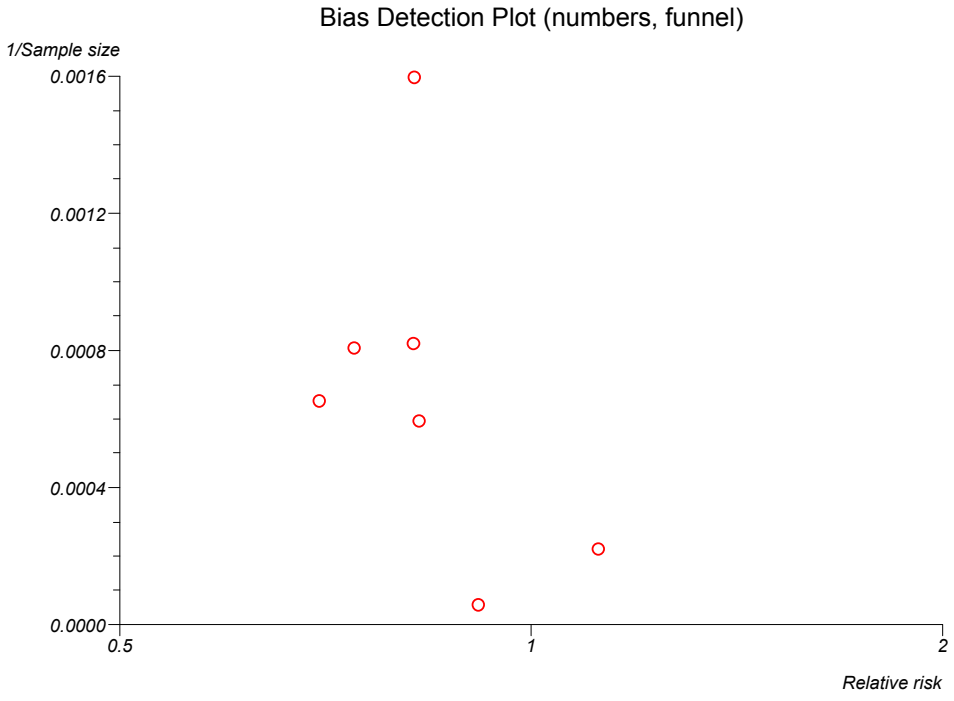
Chi-square (for pooled relative risk) = 10.809386 (df = 1) P = 0.001

Q ("non-combinability" for relative risk) = 9.928487 (df = 6) P = 0.1277

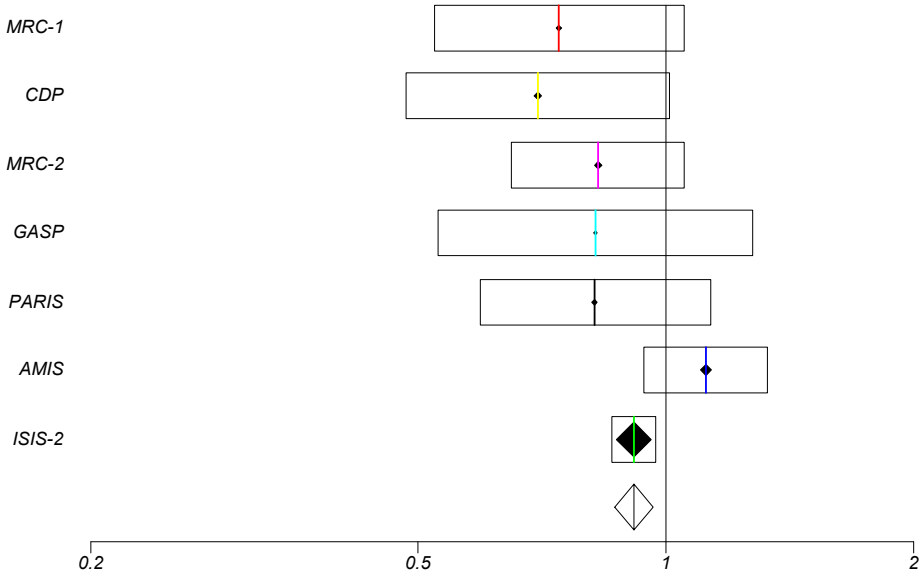
DerSimonian-Laird pooled relative risk = 0.892922
Approximate 95% CI = 0.800632 to 0.995851

DerSimonian-Laird chi-square = 4.139819 (df = 1) P = 0.0419

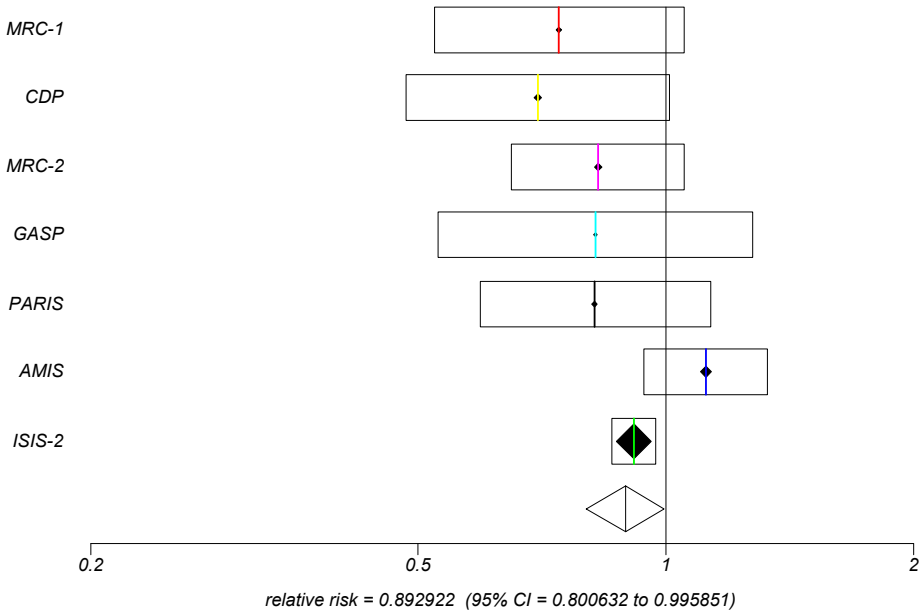

Bias Detection Plot (weights)
1/Weight
0.08
0.06
0.04
0.02
0.00
0.5
1
2
Relative risk


Bias Detection Plot (numbers, funnel)
1/Sample size
0.0016
0.0012
0.0008
0.0004
0.0000
0.5
1
2
Relative risk

Cochrane relative risk plot (fixed effects)



Cochrane relative risk plot (random effects)



Risk difference meta-analysis

Data are from Fleiss and Gross (1991) and are provided in the "Exposed total", "Exposed cases", "Non-exposed total", "Non-exposed cases" and "Study" columns of the "test" workbook.

Risk difference meta-analysis

Stratum	Table (a, b, c, d)				
1	49	67	566	557	MRC-1
2	44	64	714	707	CDP
3	102	126	730	724	MRC-2
4	32	38	285	271	GASP
5	85	52	725	354	PARIS
6	246	219	2021	2038	AMIS
7	1570	1720	7017	6880	ISIS-2

Stratum	Risk difference	95% CI (near exact)		M-H Weight	
1	-0.027697	-0.060615	0.00482	3665.35131	MRC-1
2	-0.024962	-0.051149	0.000744	5852.68972	CDP
3	-0.025639	-0.058485	0.007133	3599.329653	MRC-2
4	-0.022031	-0.072572	0.027778	1573.962937	GASP
5	-0.023141	-0.06406	0.013999	2557.420582	PARIS
6	0.011482	-0.006236	0.029241	12271.1118	AMIS
7	-0.017165	-0.028929	-0.005404	27774.885022	ISIS-2

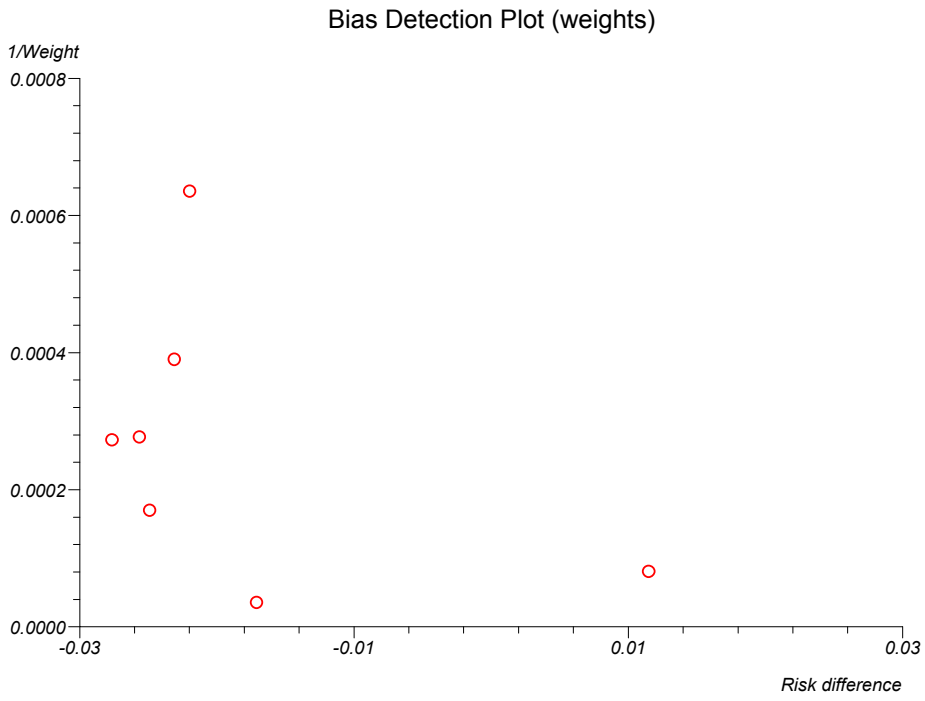
Pooled estimate (Greenland-Robins) of risk difference = -0.014263
Approximate 95% CI = -0.022765 to -0.005762

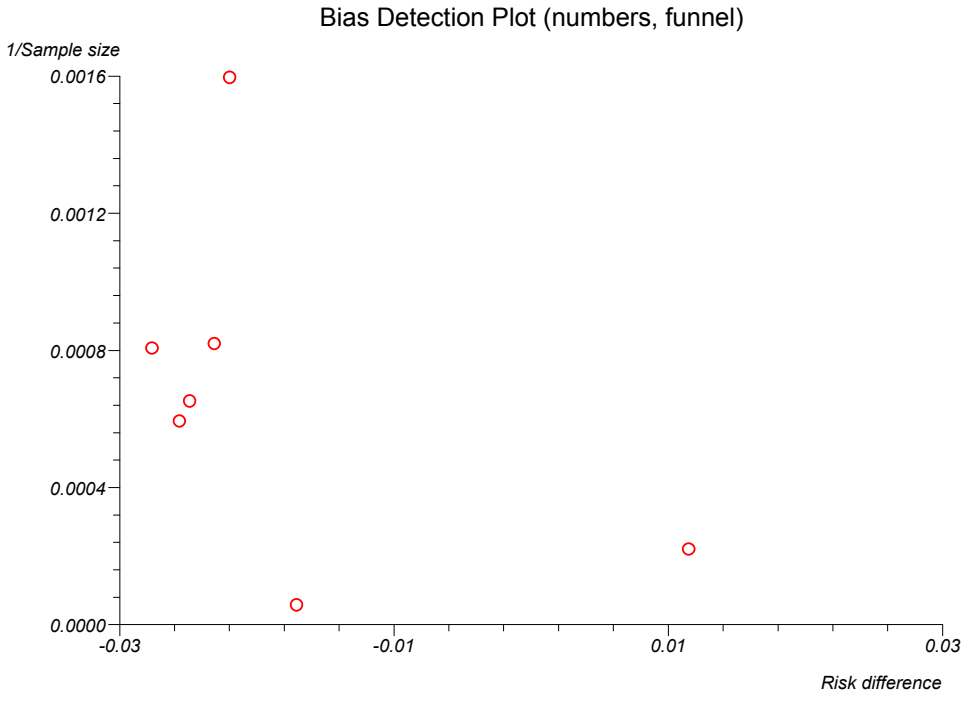
Chi-square (for pooled risk difference) = 10.812247 (df = 1) P = 0.001

Q ("non-combinability" for risk difference) = 10.461119 (df = 6) P = 0.1065

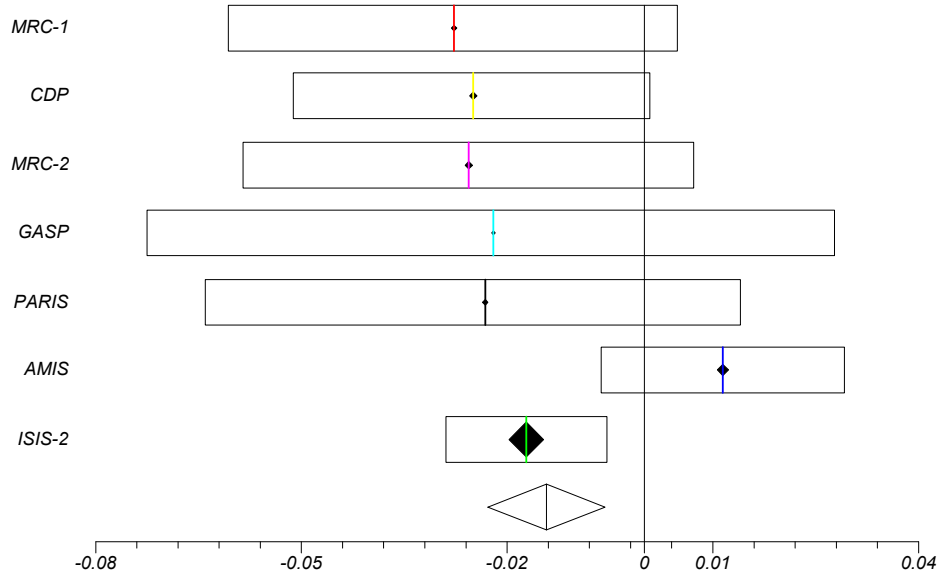
DerSimonian-Laird pooled risk difference = -0.014947
Approximate 95% CI = -0.0276 to -0.002295

DerSimonian-Laird chi-square = 5.361139 (df = 1) P = 0.0206


Bias Detection Plot (weights)
1/Weight
0.0008
0.0006
0.0004
0.0002
0.0000
-0.03
-0.01
0.01
0.03
Risk difference

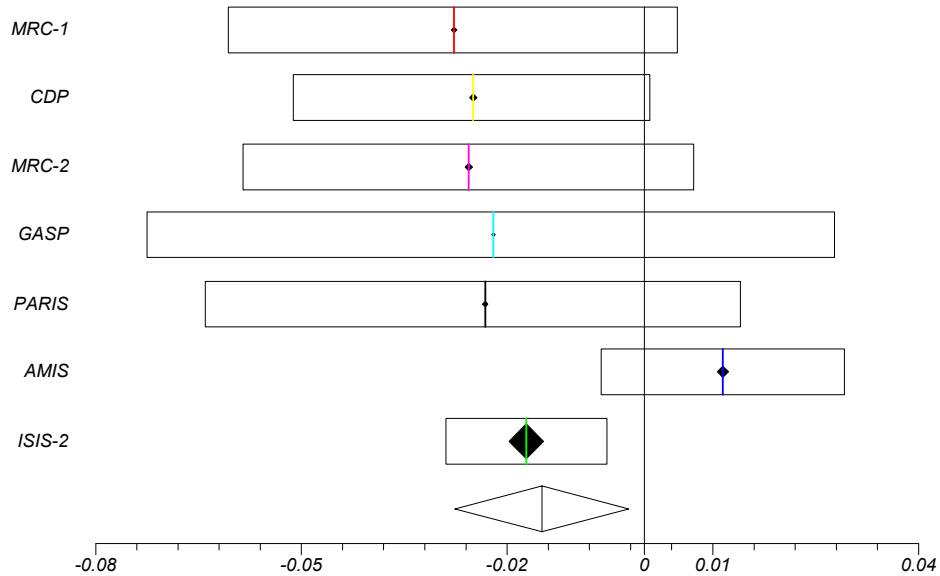

Bias Detection Plot (numbers, funnel)
1/Sample size
0.0016
0.0012
0.0008
0.0004
0.0000
-0.03
-0.01
0.01
0.03
Risk difference

Cochrane risk difference plot (fixed effects)



MH pooled risk difference = -0.014263 (95% CI = -0.022765 to -0.005762)

Cochrane risk difference plot (random effects)



DL pooled risk difference = -0.014947 (95% CI = -0.0276 to -0.002295)

Effect size meta-analysis

Data are from Freemantle (1998a) and are provided in the "Exptal. number", "Exptal. mean", "Exptal. SD", "Control number", "Control mean", "Control SD" and "Trial" columns of the "test" workbook.

Effect size meta-analysis

Stratum	J(N-2)	g	Exact 95% CI		
1	0.982018	0.826124	0.190009	1.453003	Kottkle
2	0.973875	0.171233	-0.535959	0.875877	Levinson
3	0.991545	0.564449	0.095392	1.030273	Oliver (intensive)
4	0.994034	0.350002	2.95E-17	0.698564	Oliver (standard)
5	0.973875	1.20168	0.358101	2.026698	Sulmasy
6	0.992553	0.880352	0.463946	1.29277	White
7	0.982841	0.792426	0.187013	1.389328	Wilson

Stratum	N (exptl.)	N (ctrl.)	d	Approximate 95% CI		
1	27	17	0.811269	0.18121	1.441328	Kottkle
2	16	15	0.166759	-0.538869	0.872388	Levinson
3	25	66	0.559677	0.092265	1.027089	Oliver (intensive)
4	62	66	0.347914	-0.001341	0.697169	Oliver (standard)
5	9	22	1.170286	0.341855	1.998716	Sulmasy
6	63	40	0.873796	0.459972	1.28762	White
7	23	23	0.778829	0.179356	1.378302	Wilson

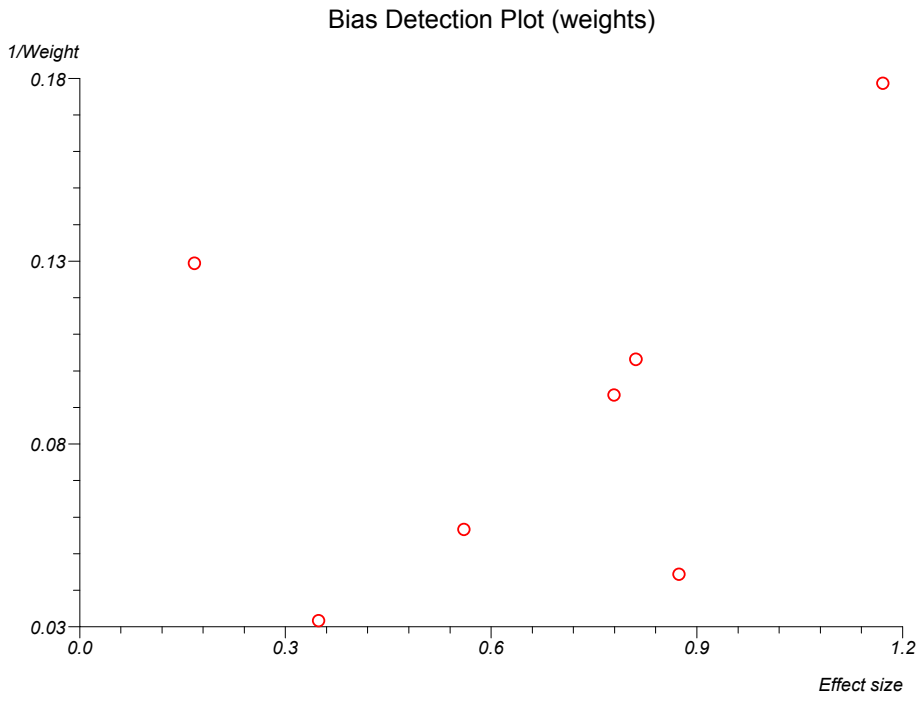
Pooled estimate of effect size d+ = 0.612354
Approximate 95% CI = 0.421251 to 0.803457

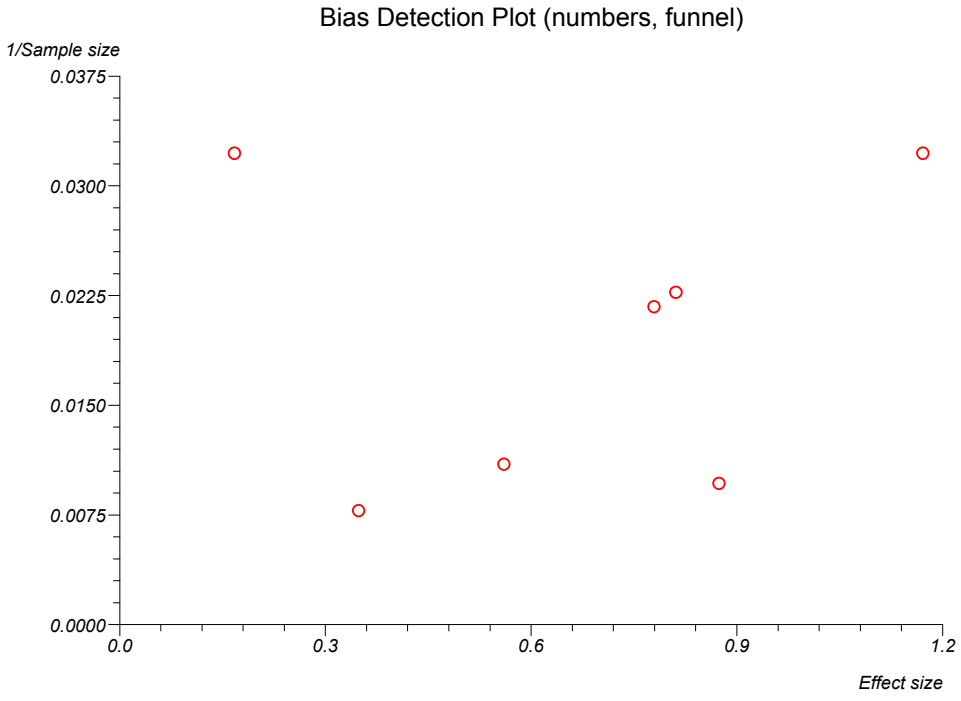
Chi-square (for d+) = 39.442584 (df = 1) P < 0.0001

Q ("non-combinability" for d+) = 7.737692 (df = 6) P = 0.258

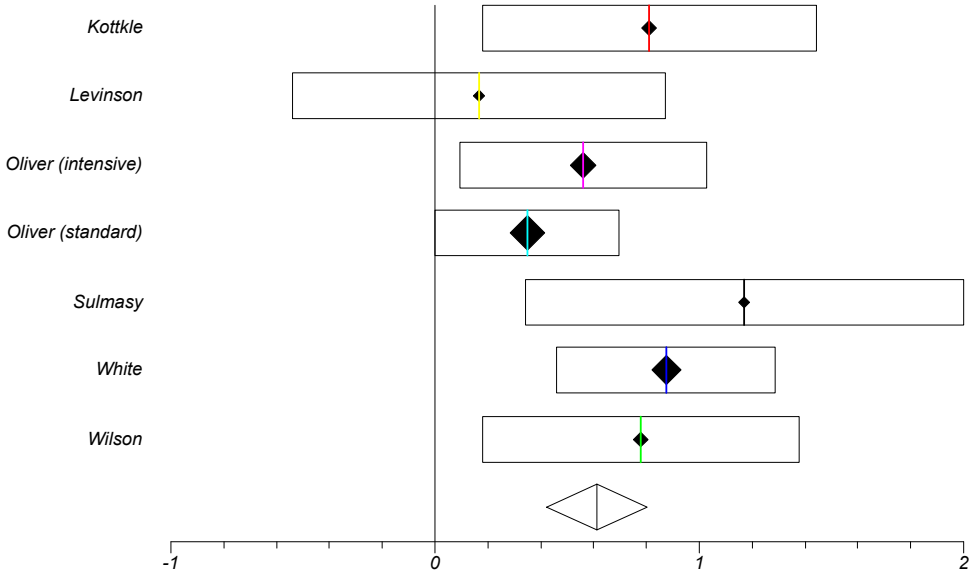
DerSimonian-Laird pooled d+ = 0.627768
Approximate 95% CI = 0.403026 to 0.85251

DerSimonian-Laird chi-square = 29.972717 (df = 1) P < 0.0001


Bias Detection Plot (weights)
1/Weight
0.18
0.13
0.08
0.03
0.0
0.3
0.6
0.9
1.2
Effect size

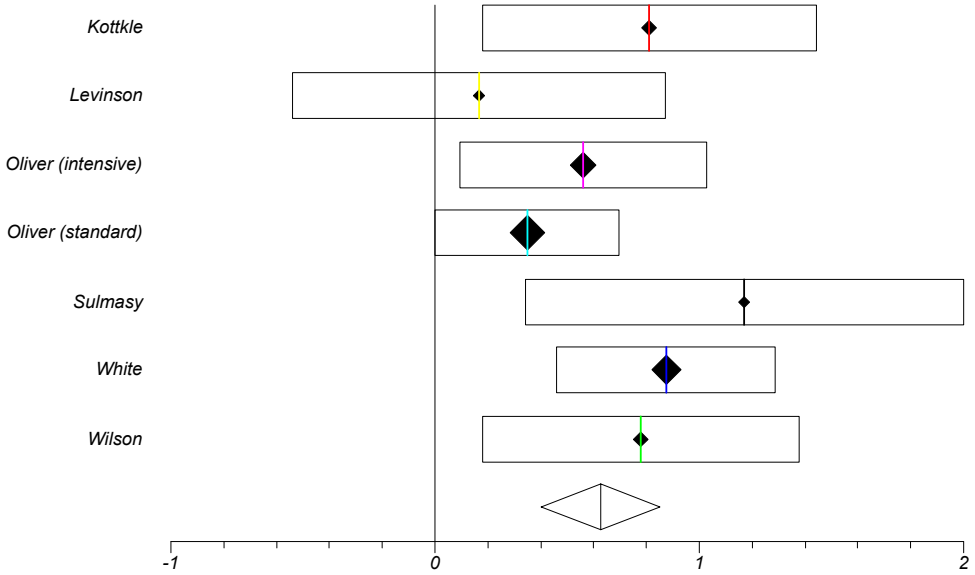

Bias Detection Plot (numbers, funnel)
1/Sample size
0.0375
0.0300
0.0225
0.0150
0.0075
0.0000
0.0
0.3
0.6
0.9
1.2
Effect size

Cochrane effect size plot (fixed effects)



pooled effect size = 0.612354 (95% CI = 0.421251 to 0.803457)

Cochrane effect size plot (random effects)



DL pooled effect size = 0.627768 (95% CI = 0.403026 to 0.85251)

Incidence rate meta-analysis

Data are from Rothman et al. (1973) and are provided in the "Exposed cases", "Exposed person-time", "Control cases", "Control person-time" and "Age split" columns of the "test" workbook.

Incidence rate difference (IRD) meta-analysis

Stratum	Table (a, person-time exposed, b, person-time not exposed)				
1	14	1516	10	1701	Age < 65
2	76	949	121	2245	Age >= 65

Stratum	IRD	95% CI (approximate)		Weight	
1	0.003356	-0.002623	0.009335	104737.09221	Age < 65
2	0.026187	0.00734	0.045034	9225.440386	Age >= 65

Pooled estimate of IRD = 0.005204
Approximate 95% CI = -0.000602 to 0.01101

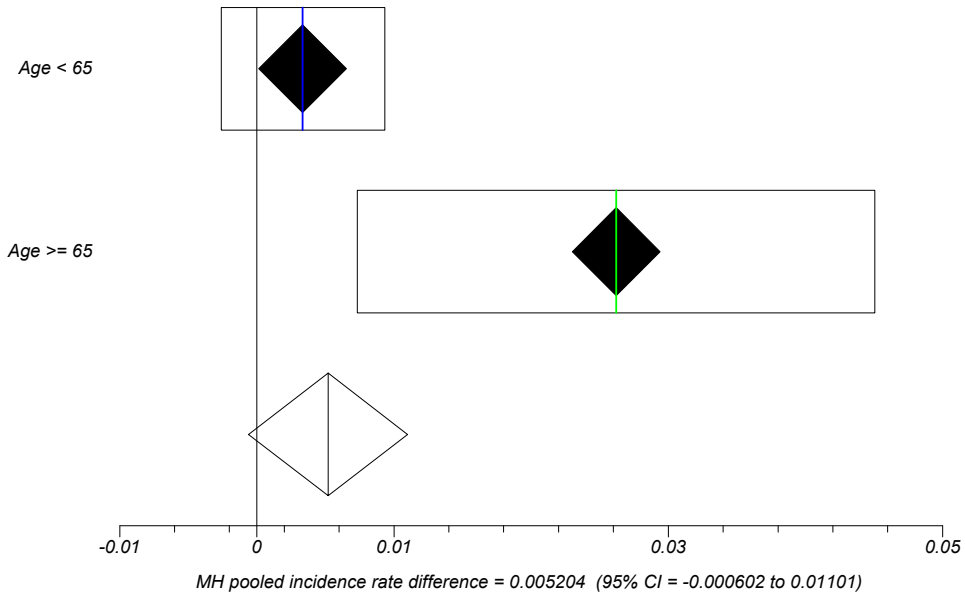
Chi-square (for pooled IRD) = 3.086435 (df = 1) P = 0.0789

Q ("non-combinability" for IRD) = 4.419452 (df = 1) P = 0.0355

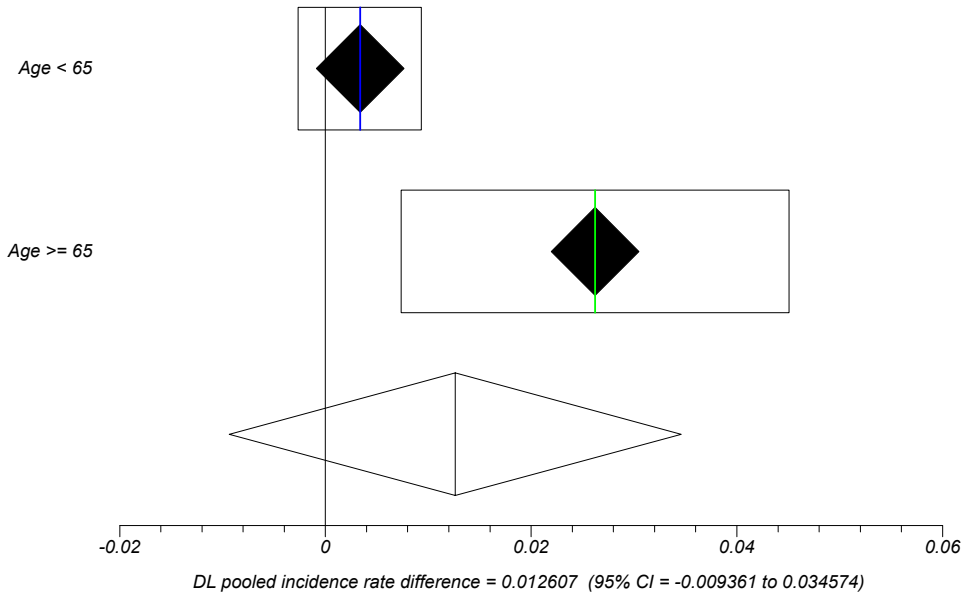
DerSimonian-Laird pooled IRD = 0.012607
Approximate 95% CI = -0.009361 to 0.034574

DerSimonian-Laird chi-square = 1.265072 (df = 1) P = 0.2607

Cochrane incidence rate difference plot (fixed effects)



Cochrane incidence rate difference plot (random effects)



Incidence rate ratio (IRR) meta-analysis

Stratum	Table (a, person-time exposed, b, person-time not exposed)				
1	14	1516	10	1701	Age < 65
2	76	949	121	2245	Age >= 65

Stratum	IRR	95% CI (exact)		Weight	
1	1.570844	0.648937	3.952809	5.833333	Age < 65
2	1.485862	1.100305	1.99584	46.680203	Age >= 65

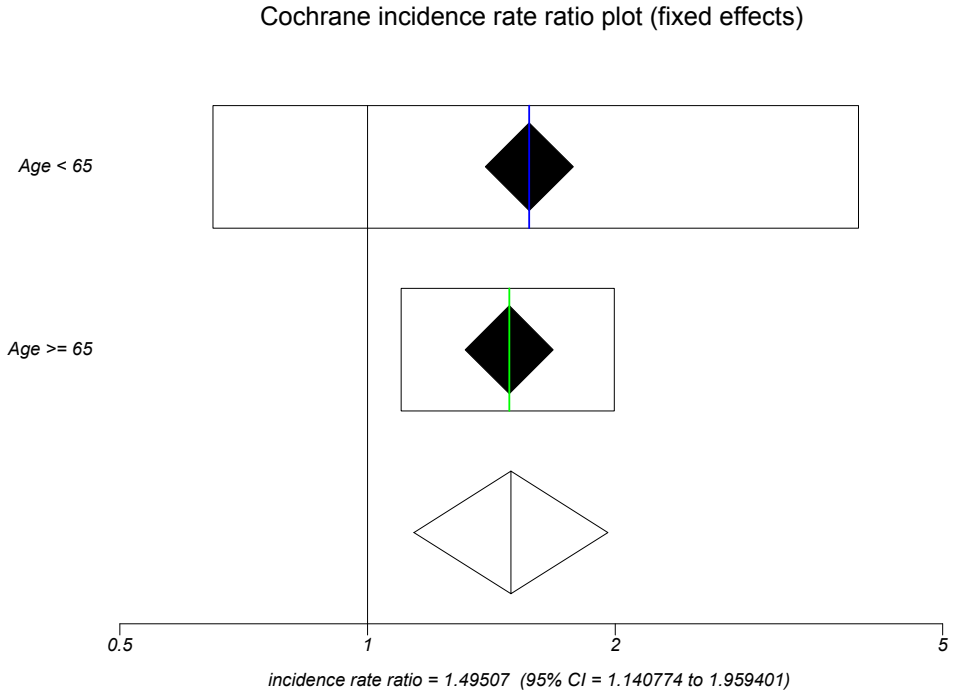
Pooled estimate of IRR = 1.49507
Approximate 95% CI = 1.140774 to 1.959401

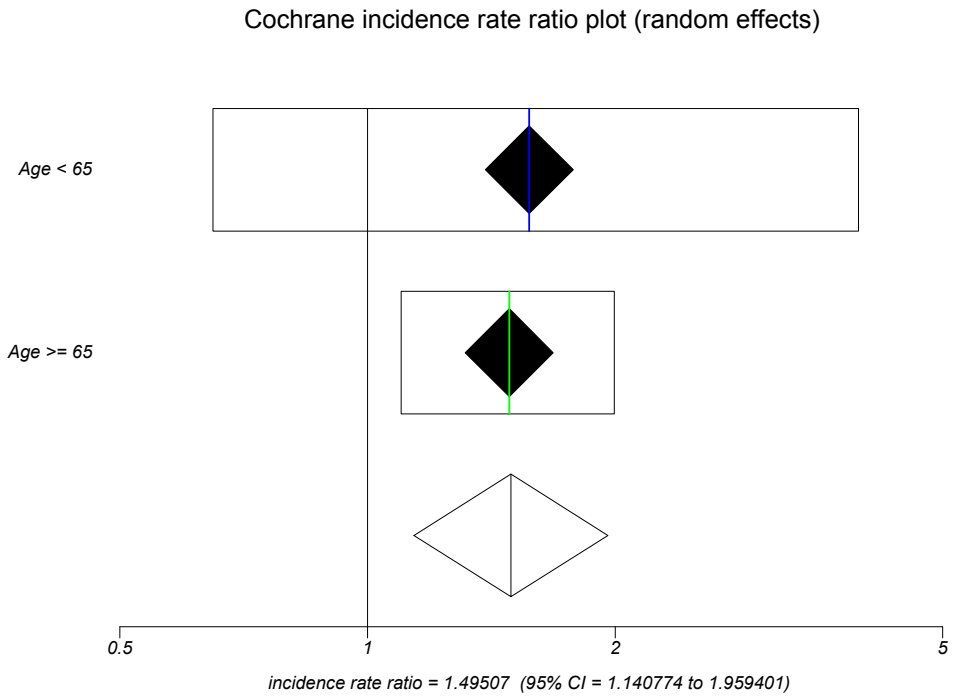
Chi-square (for pooled IRR) = 8.493705 (df = 1) P = 0.0036

Q ("non-combinability" for IRR) = 0.01604 (df = 1) P = 0.8992

DerSimonian-Laird pooled IRR = 1.49507
Approximate 95% CI = 1.140774 to 1.959401

DerSimonian-Laird chi-square = 8.493705 (df = 1) P = 0.0036




Cochrane incidence rate ratio plot (random effects)
Age < 65
Age >= 65
0.5
1
2
5
incidence rate ratio = 1.49507 (95% CI = 1.140774 to 1.959401)

Crosstabs

Data are from Armitage and Berry (1994) and are provided in the "Grief" and "Support" columns of the "test" workbook.

Crosstabs

	Support		
Grief	1	2	3
1	17	9	8
2	6	5	1
3	3	5	4
4	1	2	5

Chi-square test (r by c)

Observed	17	9	8	34	1
Expected	13.909091	10.818182	9.272727		
DChi²	0.686869	0.305577	0.174688		
Observed	6	5	1	12	2
Expected	4.909091	3.818182	3.272727		
DChi²	0.242424	0.365801	1.578283		
Observed	3	5	4	12	3
Expected	4.909091	3.818182	3.272727		
DChi²	0.742424	0.365801	0.161616		
Observed	1	2	5	8	4
Expected	3.272727	2.545455	2.181818		
DChi²	1.578283	0.116883	3.640152		
Total	27	21	18	66	
Score	1	2	3		

TOTAL number of cells = 12

WARNING: 9 out of 12 cells have EXPECTATION < 5

INDEPENDENCE

Chi-square = 9.9588 DF = 6 P = 0.1264

G-square = 10.186039 DF = 6 P = 0.117

ANOVA

Chi-square for equality of mean column scores = 5.696401

DF = 2 P = 0.0579

LINEAR TREND

Sample correlation (r) = 0.295083

Chi-square for linear trend (M^2) = 5.6598

DF = 1 P = 0.0174

COEFFICIENTS OF CONTINGENCY

Pearson's = 0.362088

Cramér's = 0.274673

Frequencies

Data are from Altman (1991) and are provided in the "IgM" column of the "test" workbook.

Frequencies

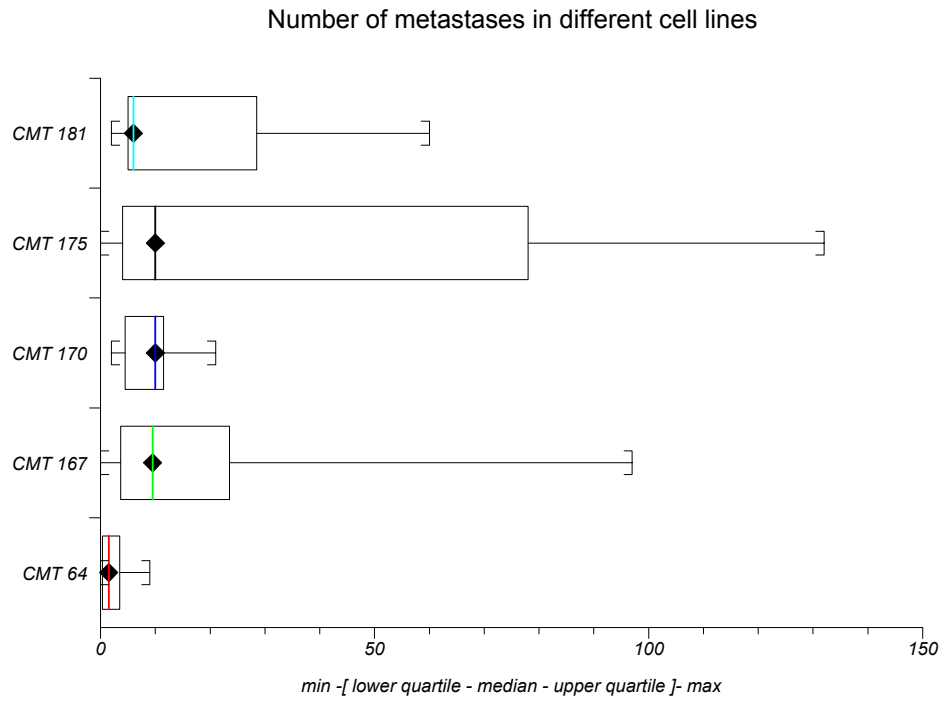
Frequency analysis for IgM:

Total = 298

Value	Frequency	Relative %	Cumulative	Cumulative Relative %
0.1	3	1.006711	3	1.006711
0.2	7	2.348993	10	3.355705
0.3	19	6.375839	29	9.731544
0.4	27	9.060403	56	18.791946
0.5	32	10.738255	88	29.530201
0.6	35	11.744966	123	41.275168
0.7	38	12.751678	161	54.026846
0.8	38	12.751678	199	66.778523
0.9	22	7.38255	221	74.161074
1	16	5.369128	237	79.530201
1.1	16	5.369128	253	84.899329
1.2	6	2.013423	259	86.912752
1.3	7	2.348993	266	89.261745
1.4	9	3.020134	275	92.281879
1.5	6	2.013423	281	94.295302
1.6	2	0.671141	283	94.966443
1.7	3	1.006711	286	95.973154
1.8	3	1.006711	289	96.979866
2	3	1.006711	292	97.986577
2.1	2	0.671141	294	98.657718
2.2	1	0.33557	295	98.993289
2.5	1	0.33557	296	99.328859
2.7	1	0.33557	297	99.66443
4.5	1	0.33557	298	100

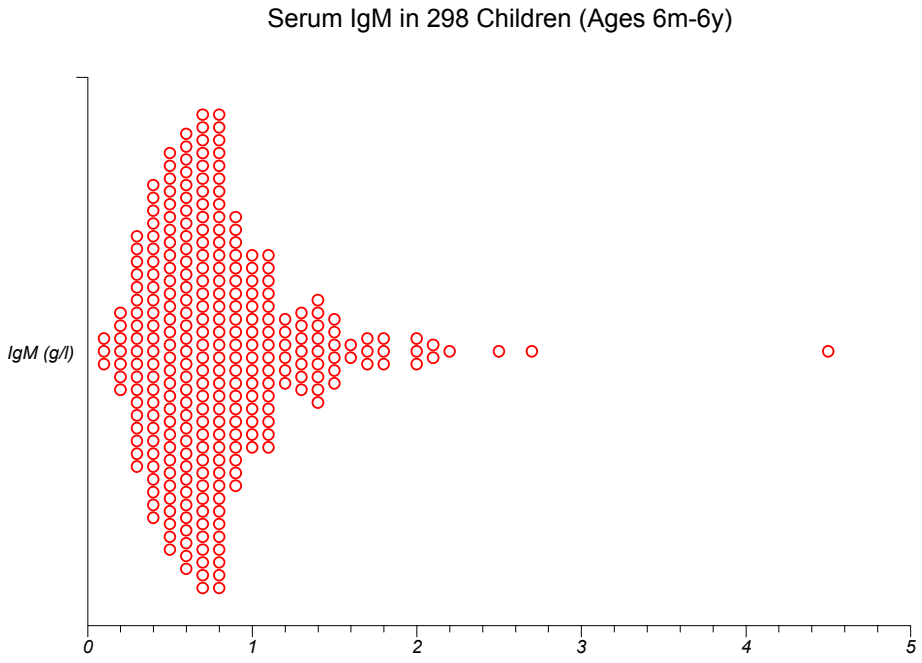
Box and whisker plot

Data are from Cuzick (1985) and are provided in the "CMT 64", "CMT 167", "CMT 170", "CMT 175" and "CMT 181" columns of the "test" workbook.



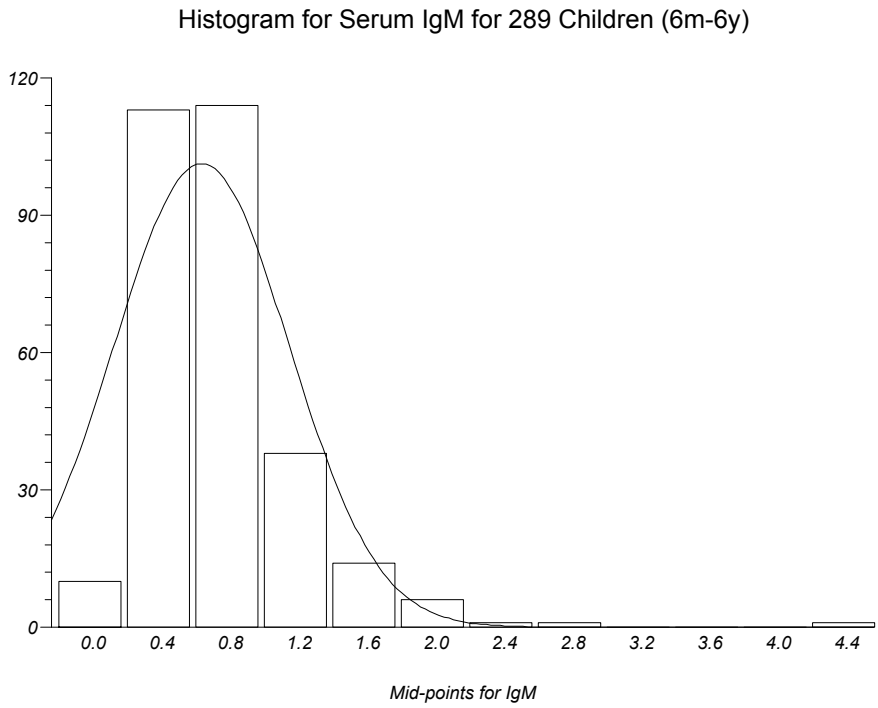
Spread Plot

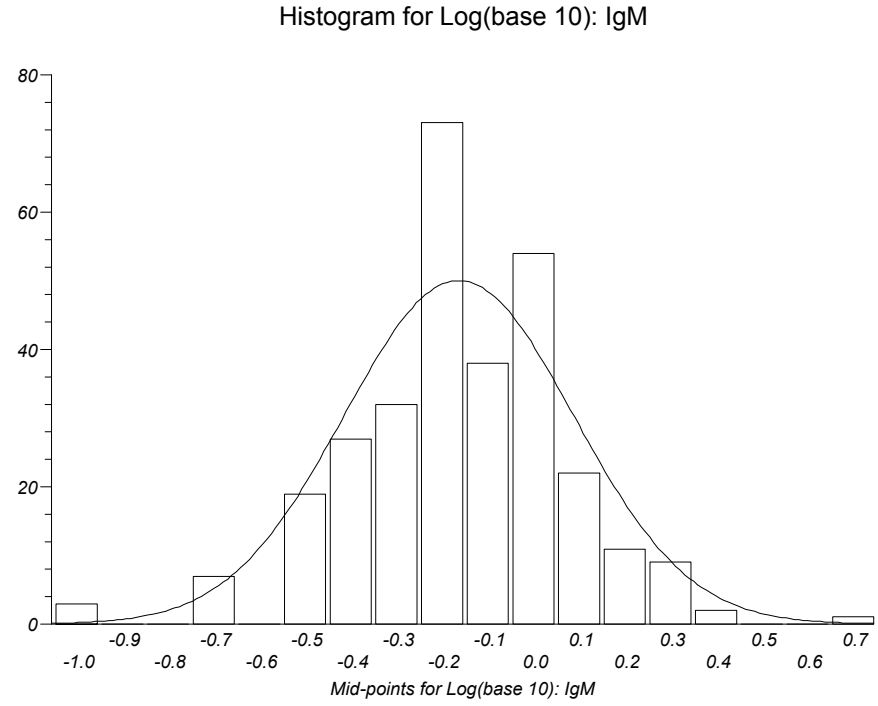
Data are from Cuzick (1985) and Altman (1991), they are provided in the "CMT 64", "CMT 167", "CMT 170", "CMT 175", "CMT 181" and "IgM" columns of the "test" workbook.



Histogram

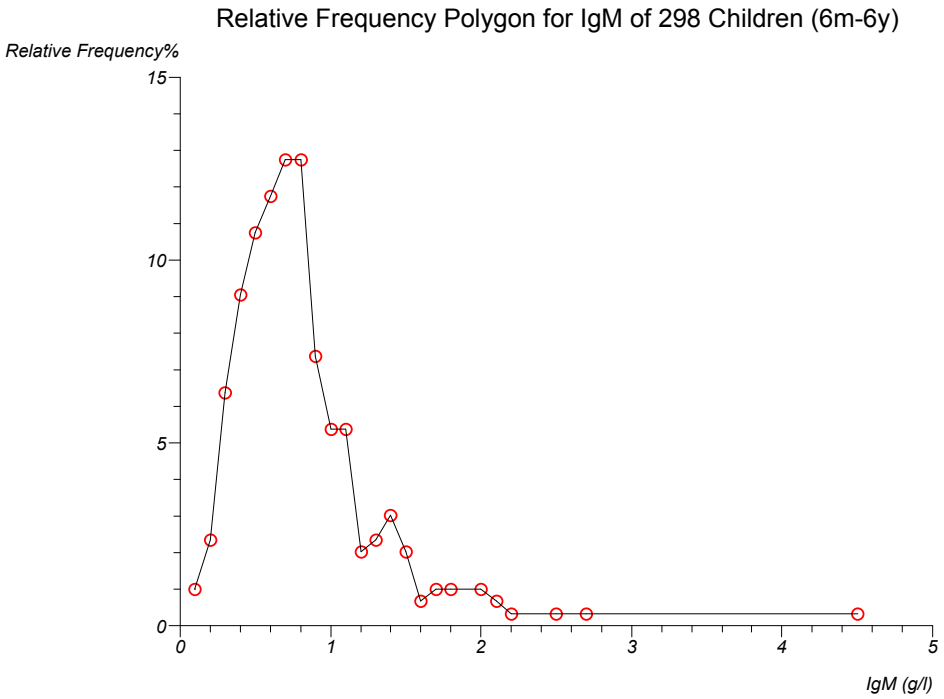
Data are from Altman (1991) and are provided in the "IgM" and "Log(base 10): IgM" columns of the "test" workbook.

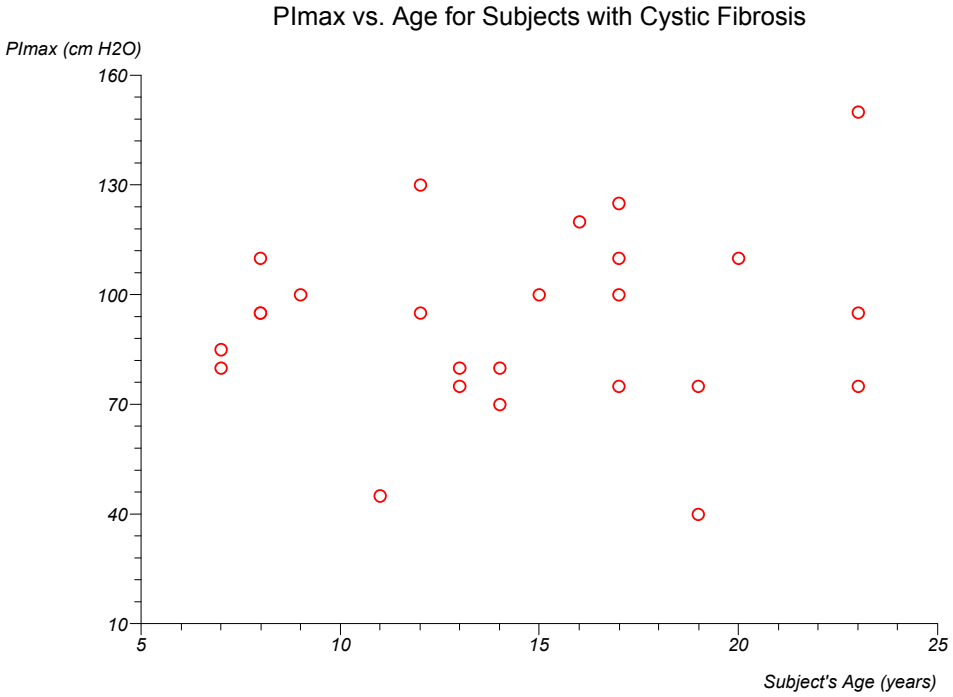




Scatter Plot

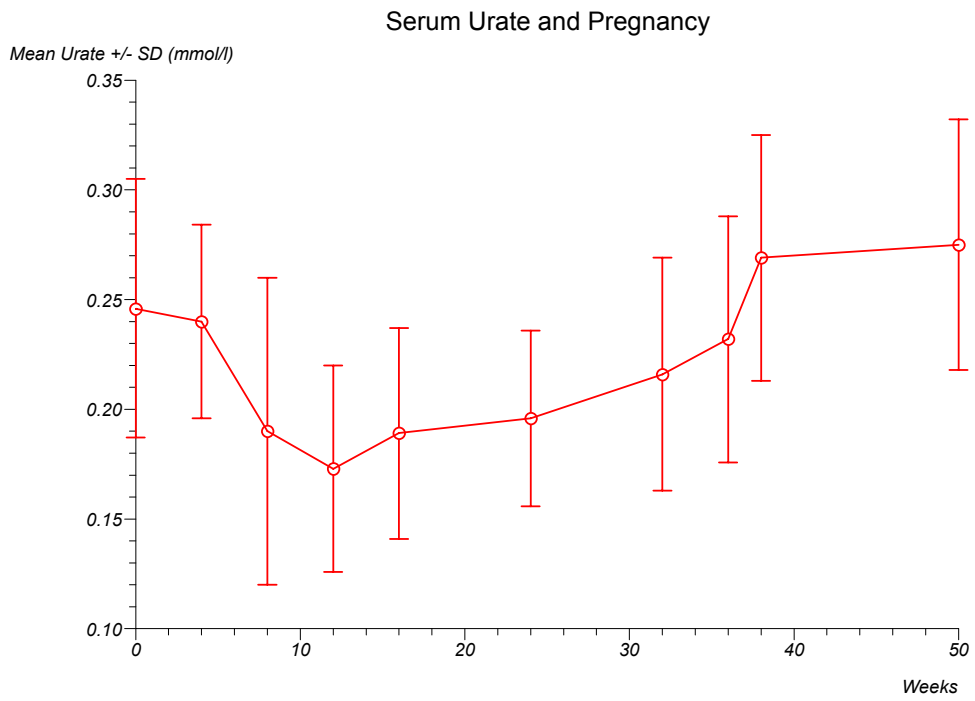
Data are from Altman (1991) and are provided in the "PImax (cm H2O)", "Subject's Age (years)", "Relative Frequency%" and "IgM Values" columns of the "test" workbook.





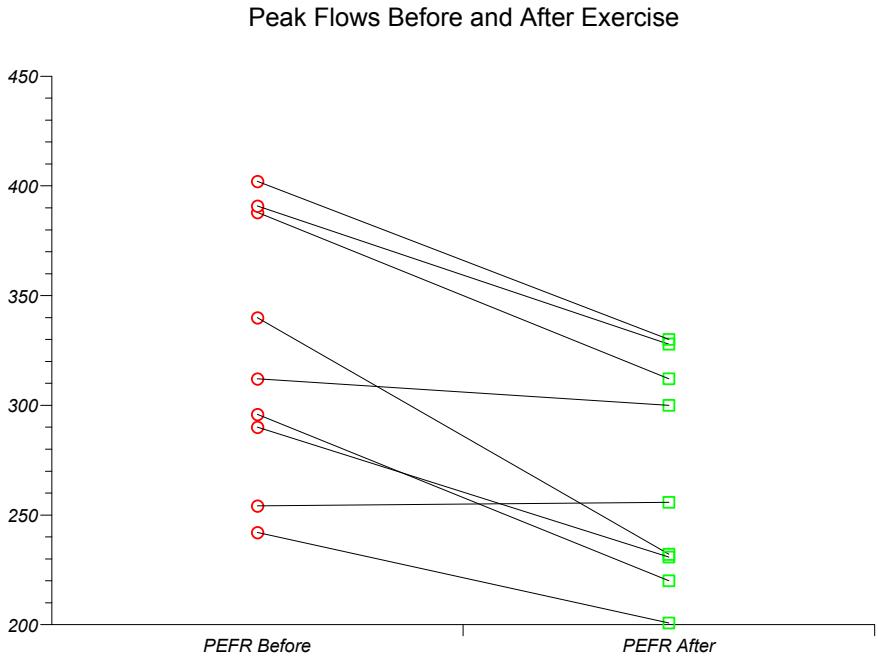
Error bar plot

Data are from Altman (1991) and are provided in the "Mean Urate (mmol/l)", "SD Urate (mmol/l)" and "Weeks since conception" columns of the "test" workbook.



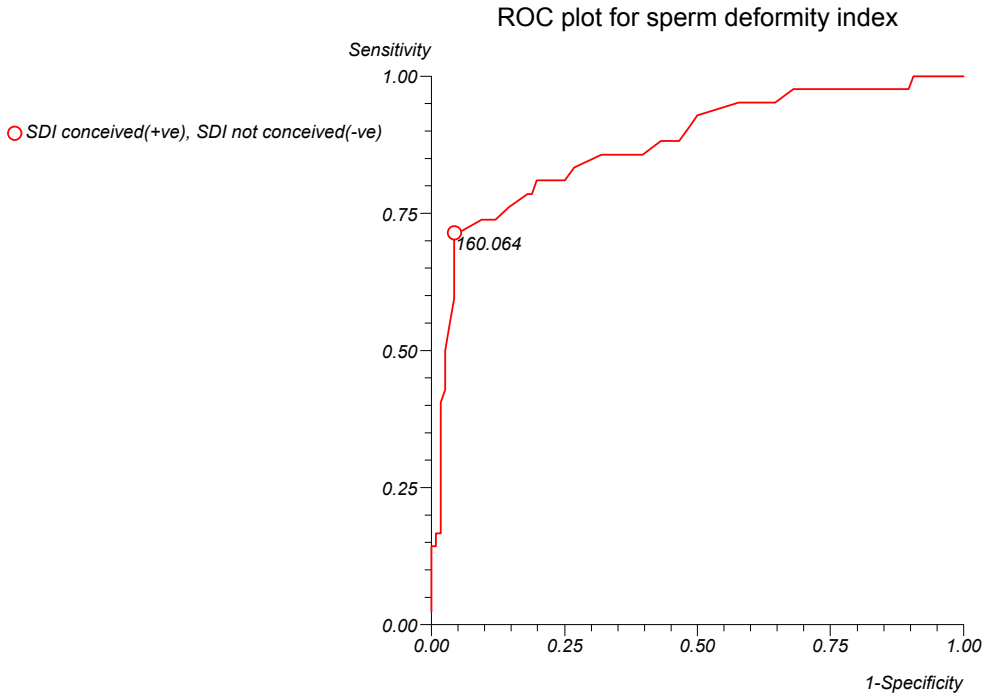
Ladder plot

Data are from Bland (1996) and are provided in the "PEFR Before" and "PEFR After" columns of the "test" workbook.



Receiver operating characteristic curve

Data are from Aziz et al. (1996) and are provided in the "SDI conceived" and "SDI not conceived" columns of the "test" workbook.



ROC Analysis

Data set: SDI conceived(+ve), SDI not conceived(-ve)

Area under ROC curve by extended trapezoidal rule = 0.875411
Wilcoxon estimate (95% CI) of area under ROC curve = 0.875411 (0.799283 to 0.951538)

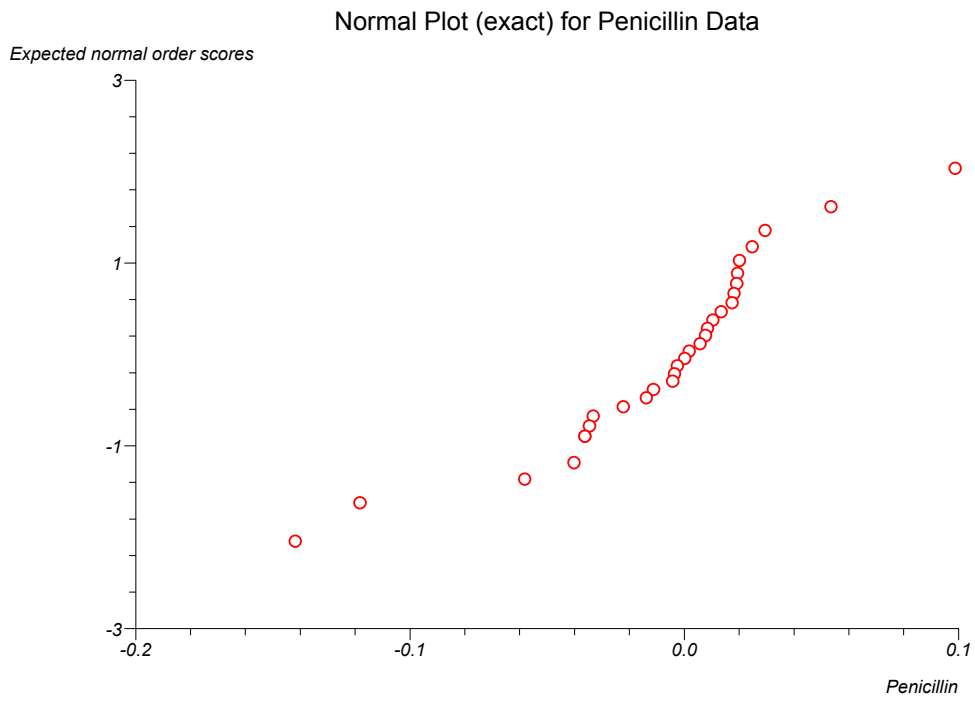
Optimum cut-off point selected = 160.064

Table at cut-off:	a	b
	30	5
	c	d
	12	111

sensitivity = 0.714286
specificity = 0.956897

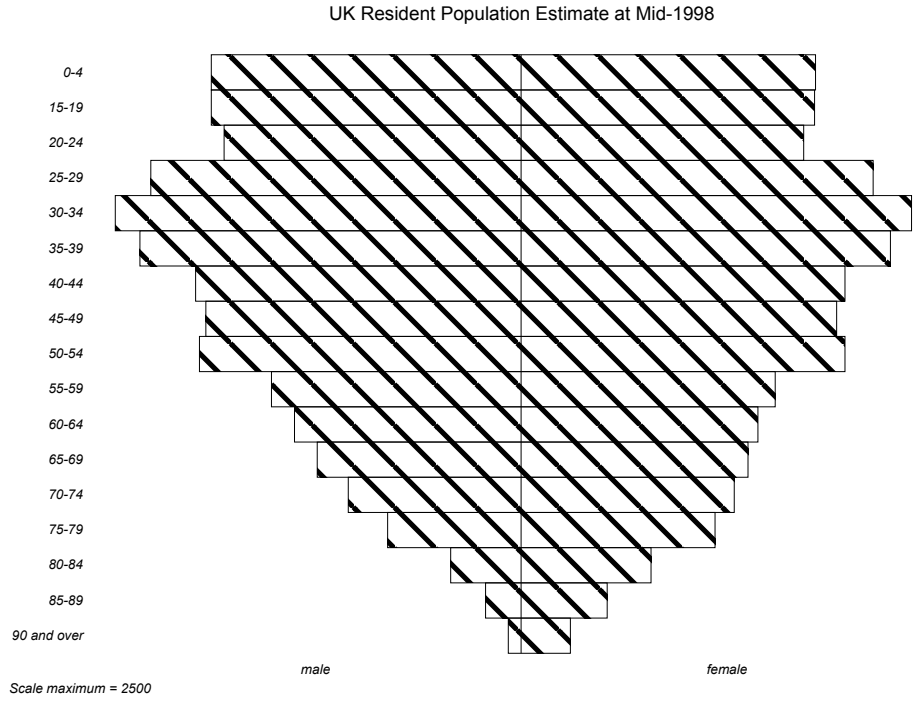
Normal Plot

Data are from Shapiro and Wilk (1965) and are provided in the "Penicillin" column of the "test" workbook.



Population Pyramid

Data are from Statbase (2000) and are provided in the "UK Mid-98 Males", "UK Mid-98 Females" and "UK Mid-98Age Bands" columns of the "test" workbook.



Comparisons with other statistical resources

Knowledge support

The following table compares the statistical knowledge content available within the standard software of StatsDirect, SPSS (all modules), Minitab and Stata (SPSS Corporation 2000; Minitab Corporation 1999; Stata Corporation 1999):

Knowledge	StatsDirect	SPSS	Minitab	Stata
Interface	book-like WinHelp 2000	Windows Help 4	Windows Help 4	text only proprietary hypertext
Basic statistical concept topics (e.g. confidence intervals)	yes	no	no	no
Basic epidemiological concept topics (e.g. bias)	yes	no	no	no
Worked examples	full	full	part	no
References for examples	yes	no	no	na
Inferences with examples	yes	no	no	na
Medical research context	yes	no	no	no
Reference list	full	full	part	none
References for functions	full	part	part	none

Access to statistical methods

The following table compares the statistical methods available within the standard software of StatsDirect, SPSS, Minitab and Stata (SPSS Corporation 2000; Minitab Corporation 1999; Stata Corporation 1999):

Goals	Methods	StatsDirect	SPSS	Minitab	Stata
Describing data	Descriptive statistics	yes	yes	yes	yes
	Frequencies	yes	yes	yes	yes
	Cross tabulation	yes	yes	yes	yes
	Box & whisker plot	yes	yes	yes	yes
	Spread plot	yes	no	ungrouped dotplot	atypical dotplot

Goals	Methods	StatsDirect	SPSS	Minitab	Stata
	Frequency histogram	yes	yes	yes	yes
	Normal plot	yes	yes	yes	yes
Transforming data	Ranking	yes	yes	yes	yes
	Sorting	yes	yes	yes	yes
	Normal scores	yes	yes	indirect	yes
	Free-form arithmetic	yes	yes	yes	yes
	Logarithms	yes	yes	yes	yes
	Logit	yes	indirect	indirect	indirect
	Probit	yes	indirect	no	no
	Angle	yes	indirect	indirect	indirect
	Cumulative	yes	indirect	no	yes
	Ladder of powers	yes	no	no	yes
	Box-Cox	no	no	yes	yes
	Standardisation	yes	partial	yes	indirect
	Matrix tools	transpose only (rotate)	yes	yes	yes
	Dummy variables	yes	indirect	indirect	indirect
	Split data into groups given grouping variable	yes	indirect	yes	yes
	Combine groups and generate grouping variable	yes	indirect	yes	yes
	Comprehensive spreadsheet functions	yes	no	no	no
	Data search and replace interface	yes	no	no	no
	Definable translation between text and numeric data	yes	yes	yes	yes
	Intervals from date and time data	yes	indirect	indirect	indirect
	Pairwise differences	yes	no	yes	no
	Pairwise means	yes	no	yes	no
	Pairwise slopes	yes	no	yes	no

Goals	Methods	StatsDirect	SPSS	Minitab	Stata
Theoretical distributions	Gaussian normal (and log-normal by transformation)	yes	yes	yes	yes
	chi-square	yes	yes	yes	yes
	Student t	yes	yes	yes	yes
	F (variance ratio)	yes	yes	yes	yes
	Studentized range	yes	no	yes	no
	binomial	yes	yes	yes	yes
	Poisson	yes	yes	yes	no
	Kendall	yes	no	no	no
	Spearman/Hotelling-Pabst	yes	no	no	no
	non-central Student t	yes	partial	no	no
	uniform random numbers	yes	yes	yes	yes
	normal random numbers	yes	yes	yes	yes
	binomial random numbers	yes	yes	yes	no
	Poisson random numbers	yes	yes	yes	no
	gamma random numbers	yes	yes	yes	no
	exponential random numbers	yes	yes	yes	no
Designing studies	Randomization	yes	indirect	indirect	indirect
	Sample sizes	yes	no	yes	yes
Inference from a single group	Sign test	yes	partial	indirect	yes
	Single proportion	yes	indirect	yes	yes
	Single sample t test	yes	yes	yes	yes
	Single sample z test	yes	no	yes	indirect
	Reference range	yes	no	no	no
	Quantile with confidence interval	yes	no	median only	yes
	Mean with confidence interval	yes	yes	yes	yes

Goals	Methods	StatsDirect	SPSS	Minitab	Stata
	Poisson confidence interval	yes	no	no	yes
	Shapiro-Wilk W test for non-normality	yes	yes	similar	similar
	Chi-square goodness of fit test	yes	yes	no	no
Comparison of two independent groups	Unpaired t test	yes	yes	yes	yes
	Unpaired z test	yes	no	yes	no
	F (variance ratio) test	yes	indirect	indirect	yes
	Mann-Whitney test	yes	yes	yes	yes
	Smirnov two sample test	yes	yes	indirect	yes
	Chi-square 2 by 2 test	yes	indirect	yes	yes
	Fisher's exact test	yes	indirect	no	yes
	Two independent proportions	yes	no	partial	yes
	Crossover ANOVA	yes	no	no	no
	Chi-square 2 by k test	yes	indirect	yes	yes
	Equality of variance	yes	yes	yes	indirect
	Agreement analysis and plots	yes	no	no	no
	ROC curves	yes	no	no	indirect
	Box & whisker plots	yes	yes	yes	yes
	Spread plots	yes	no	ungrouped dotplot	atypical dotplot
Comparison of a pair of groups	Paired t test	yes	yes	yes	yes
	Pared z test	via single sample test	no	yes	no
	Wilcoxon signed ranks test	yes	yes	via single sample test	yes
	McNemar and Liddell tests	yes	partial indirect	no	partial
	Maxwell's test	yes	no	no	yes
	Paired proportions	yes	no	no	no

Goals	Methods	StatsDirect	SPSS	Minitab	Stata
	Ladder plots	yes	no	no	no
Comparison of more than two groups	One way ANOVA	yes	yes	yes	yes
	Tukey multiple contrasts	yes	yes	yes	indirect
	Scheffé multiple contrasts	yes	yes	no	yes
	Dunnett multiple contrasts with a control group	yes	yes	yes	no
	Bonferroni multiple contrasts	yes	yes	no	yes
	Newman-Keuls multiple contrasts	yes	yes	no	no
	Two way randomized block ANOVA	yes	indirect	yes	indirect
	Two way randomized block ANOVA with repeated observations	yes	indirect	yes	indirect
	Fully nested random ANOVA	yes	indirect	yes	indirect
	User defined ANOVA by generalised linear model	no	yes	yes	yes
	User defined ANOVA for a balanced design	no	indirect	yes	yes
	Kruskal Wallis test	yes	yes	yes	yes
	Friedman test	yes	yes	yes	yes
	Cuzick's test for trend	yes	no	no	yes
	Latin square ANOVA	yes	no	no	no

Goals	Methods	StatsDirect	SPSS	Minitab	Stata
	Agreement analysis and plots	yes	partial indirect	no	partial
	Equality of variance	yes	yes	yes	yes
	Kappa and Scott inter-rater agreement	yes	partial	no	partial
	Covariance analysis	yes	indirect	partial	yes
	r by c chi-square and G-square tests	yes	indirect	partial	yes
	Box & whisker plots	yes	yes	yes	yes
	Spread plots	yes	no	ungrouped dotplot	atypical dotplot
Relationship between groups	Simple linear regression and correlation	yes	yes	yes	yes
	Nonparametric linear regression	yes	no	no	indirect
	General linear regression	yes	yes	yes	yes
	Linearity tests	yes	yes	no	no
	Covariance analysis	yes	indirect	partial	yes
	Polynomial regression	yes	indirect	partial	yes
	Linearized estimates for geometric, hyperbolic and exponential curves.	yes	partial	no	no
	General binary logistic regression	yes	yes	yes	yes
	General ordinal logistic regression	no	yes	yes	yes
	Generalised linear models for user defined functions	no	yes	yes	yes
	Poisson regression	no	yes	no	yes
	Probit analysis	yes	yes	yes	yes
	discriminant analysis	yes	yes	yes	yes

Goals	Methods	StatsDirect	SPSS	Minitab	Stata
	Principal components analysis	yes	indirect	yes	yes
	Factor analysis	no	yes	yes	yes
	Cluster analysis	no	yes	yes	partial
	Multivariate ANOVA	no	yes	yes	partial
	r by c contingency table tests	yes	indirect	partial	yes
	Kendall's rank correlation	yes	yes	yes	yes
	Spearman's rank correlation	yes	yes	yes	yes
	Scatter plots	yes	yes	yes	yes
	Regression function plots with intervals	yes	partial	partial	yes
	Residual plots from regressions	yes	yes	indirect	yes
Specific Epidemiological methods	Prospective risk analysis	yes	no	no	partial
	Retrospective risk analysis	yes	no	no	partial
	Number needed to treat	yes	no	no	no
	Standardized mortality ratios	yes	no	no	yes
	Diagnostic test table (2 by 2) analysis	yes	no	no	no
	Likelihood ratios (k levels) with confidence intervals	yes	no	no	no
	Comparison of two incidence rates with person-time data	yes	no	no	yes
	Inference from screening test error rates	yes	no	no	no
	Population pyramids	yes	no	no	no

Goals	Methods	StatsDirect	SPSS	Minitab	Stata
Survival analysis	Kaplan-Meier estimates of survival and hazard	yes	yes	yes	yes
	Life tables	yes	yes	yes	yes
	Log-rank and Wilcoxon tests for two groups	yes	partial	partial	partial
	Log-rank and Wilcoxon tests for more than two groups (with trend)	yes	yes	partial	partial
	Wei-Lachin test	yes	no	no	no
	Cox regression	yes	yes	yes	yes
	Survival plots	yes	yes	yes	yes
Meta-analysis	Odds ratio	yes	no	no	yes
	Mantel-Haenszel test	yes	yes	no	indirect
	Woolf statistics for 2 by 2 stratified data	yes	no	no	no
	Peto odds ratio	yes	no	no	yes
	Risk ratio	yes	no	no	yes
	Risk difference	yes	no	no	yes
	Incidence rate ratio	yes	no	no	no
	Incidence rate difference	yes	no	no	no
	Effect size	yes	no	no	no
	Plots (Cochrane/"forest")	yes	no	no	yes
	Bias detection plots	yes	no	no	yes

Samples of interaction and output

Analysis of a two by two tabulation of counts and comparison of two independent groups of observations sampled from a non-normal distribution are taken as examples of statistical methods commonly used in medical research (Armitage and Berry, 1994; Altman, 1991; Bland, 1996). Interaction between user and software for these methods is examined for StatsDirect, SPSS, Minitab and Stata below.

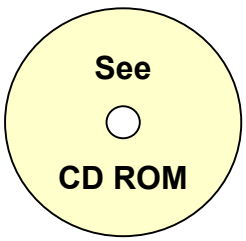
The sample counts are (Armitage and Berry, 1994):

		Outcome:	
		Dead	Alive
Treatment / Exposure:	A	41	216
	B	64	180

The sample data for two independent groups are (Conover, 1999):

Farm Boys: 14.8, 7.3, 5.6, 6.3, 9, 4.2, 10.6, 12.5, 12.9, 16.1, 11.4, 2.7.

Town Boys: 12.7, 14.2, 12.6, 2.1, 17.7, 11.8, 16.9, 7.9, 16.0, 10.6, 5.6, 5.6, 7.6, 11.3, 8.3, 6.7, 3.6, 1.0, 2.4, 6.4, 9.1, 6.7, 18.6, 3.2, 6.2, 6.1, 15.3, 10.6, 1.8, 5.9, 9.9, 10.6, 14.8, 5.0, 2.6, 4.0.



Four video (statsdirect session.avi, spss session.avi, minitab session.avi and stata session.avi) files are located on the CD ROM of the electronic part of this thesis. The video clips demonstrate interaction with StatsDirect, SPSS, Minitab and Stata for analysis of the data above.

Orientating users

This section considers aspects of software that orientate a user seeking to perform a specific statistical method or seeking to accomplish a statistical goal for which the user might be unsure of the most appropriate method.

StatsDirect orientates the user in a hierarchical grouping of statistical methods that exists both in its menu structure and in the navigation structure of its help system. Goal focused orientation is achieved through the statistical method selection item in the help menu and through the index of key words in the book-like help system.

SPSS and Minitab present menu structures with hierarchical groupings of statistical methods. Neither SPSS nor Minitab present goal focused methods of orienting the user within the range of functions available in the software. SPSS has some goal-oriented titles in its sub-menu structure.

Stata presents only a typed-in command interface to its functions; this is supported by a text-only, non-Windows hypertext help system.

Interacting with users over data

For the contingency table analysis, StatsDirect prompts the user to enter the data for the two by two contingency table directly via an interface that describes each of the constituent cells. StatsDirect also prompts the user to specify the origin of the data as a case-control study, a cohort study or neither. There is no clearly accessible way to enter summary tables into SPSS for the relevant analyses. Instead, SPSS requires the user to enter contingency table data as two variables that correspond to the two dimensions of the table that is then constructed via the "crosstabs" function. Minitab permits the user to prepare the contingency table in a worksheet and then to select the relevant columns. The Minitab help system does not describe the composition of contingency tables. In order to run analyses of contingency tables, Stata provides commands with which users type in cell

frequencies decomposed row by row. Stata provides different commands for the analysis of case-control as distinct from cohort study data.

For the two independent samples of ordinal scores above, StatsDirect prompts the user to specify the data in spreadsheet form prepared either in two separate columns, or in one column of data with a corresponding column of group/sample identifiers. SPSS accepts only data that have been prepared as single variable/column with a matching group identifier variable. Minitab accepts only data presented in two separate columns. Stata provides a command that users type in, specifying the relevant column names for data (split, or combined and linked with a group identifier).

Interaction with users over results

If the user selects a two by two chi-square test then StatsDirect provides a comparison of the proportions in the contingency table by the chi-square method and (if the user accepts a prompt) by Fisher's exact method. In addition, StatsDirect provides epidemiological risk statistics (relative risk for cohort studies and odds ratio for case-control studies) with confidence intervals if the user specifies the type of study when prompted. If the contingency table is decomposed into two indicator variables, one for row observations and the other for column observations then the "crosstabs" function in SPSS gives a range of statistics relating to different analytical goals for categorical data, but no epidemiological risk statistics. Minitab provides only a bare chi-square test for the two by two table; it does not have any entries for odds ratio, relative risk or Fisher's exact test in its help system. Stata provides risk statistics that relate to the command used (either for cohort or case-control studies).

Both StatsDirect and Minitab provide a confidence interval, that is appropriate for the comparison of two sample medians, in addition to the Mann-Whitney test results. SPSS provides an approximate P value for the Mann-Whitney test statistic and fails in its calculation of an exact P value for the example above.

Stata provides only a P value for the Kruskal-Wallis test and expects the user to know that this equates to a Mann-Whitney test when two samples are specified.

Only StatsDirect provides context-sensitive results in reports. The help system in StatsDirect assists the user with inference by presenting worked examples with inferences for each method. Minitab provides a section on interpretation of results in its help system entries for some of its function. Stata does not support inference. StatsDirect encourages users to make inference from confidence intervals over and above P values; none of SPSS, Minitab or Stata give this steer.

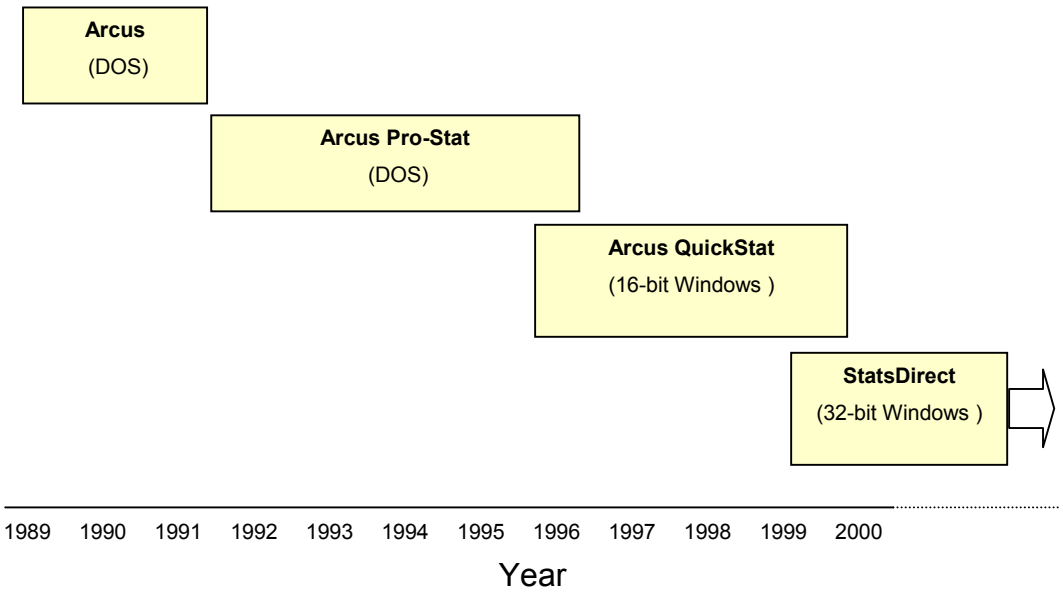
StatsDirect explains basic statistical and epidemiological concepts, such as confidence intervals, bias and confounding, to users via its help system. None of SPSS, Minitab or Stata explain these or other basic concepts to users.

Evidence of use and application

Distribution of software

Indirect evidence of application of the software from the work of this thesis comes from monitoring of the distribution of the software.

Chart of software output and distribution periods:



From 1989 to 1995, software for DOS platforms, known as Arcus ProStat, was distributed as a shareware. The distribution of Arcus ProStat was not monitored but there were 634 voluntary registration returns.

From 1997 to 2000, software for 16-bit Windows platforms, known as Arcus QuickStat (Biomedical), was distributed by the Publisher Addison Wesley Longman Ltd.. Downloads of updates to the software from the author's web site were monitored; people from 1287 different email addresses downloaded an update.

From 1999 onwards, software for 32-bit Windows platforms, known as StatsDirect, has been distributed directly by the author via the web site www.statsdirect.com. People from 4334 different email addresses downloaded the software in the period from 1st June 1999 to 31st October 2000.

Citations and reviews

The statistical use of software from the work of this thesis has been cited in peer reviewed scientific papers and books; examples that the author has examined are Bundred et al. (2000a, 2000b), Hesse (2000), Onion et al. (1996), Aziz et al. (1996), Pirmohamed et al. (1996), Williams et al. (1996), Dowrick and Buchan (1995), Wilson et al. (1995), Savage et al. (1995) and Armitage and Berry (1994).

Software from the work of this thesis has been reviewed independently of the author. All reviews have strongly recommended use of the software, and, where a rating scale has been quoted, the reviewers have awarded either top or next to top rating (Mitchell, 2000; Fitzpatrick, 2000; Freemantle, 1998a, 2000; Elliott, 1998; Sithers, 1997; Honeyball, 1997; Lee, 1995; Freeman 1995a, 1995b; Jenkins 1995; Lordan 1995; Sellu, 1994; Adler, 1994).

Discussion and Conclusions

Evaluation of developments against original aims

"A habit of basing convictions upon evidence, and of giving to them only that degree or certainty which the evidence warrants, would, if it became general, cure most of the ills from which the world suffers."
Russell, Bertrand (1872-1970)

This section examines the development of the software for this thesis towards the broad aim of building substantial potential in a software resource to improve statistical appreciation and practice in medical research.

Statistical computer software has been criticised for giving its users false confidence when performing analyses and thereby distancing relevant statistical appreciation from common research practice (Altman, 1994). Commonly used software packages that are promoted as statistical tools for medical researchers, such SPSS, Minitab and Stata, do not, unlike the software of this thesis, comprehensively provide explanations of basic statistical concepts linked with the help content attached to most of their statistical functions (SPSS Corporation, 2000; Minitab Corporation 1999; Stata Corporation, 1999).

A key development of the software for this thesis was the efficient linkage of statistical knowledge with descriptions of the operation of the functions of the software. For this purpose, statistical knowledge was arranged hierarchically; for example, detailed explanations of statistical concepts were hyperlinked beneath more applied topics in a hypertext navigation tree. Some basic epidemiological and statistical principles were arranged into discrete sections (e.g. bias) that were listed independently in the index of the help system. The content and organisation of this statistical knowledge support system was moulded by opportunistic feedback, over a period of eleven years, from researchers, students and statistician educators using the software of this thesis. In this way, statistical theory was brought closer to practice.

The statistical knowledge support provided by the software of this thesis could be further improved by eliciting more feedback, particularly from educational settings where the use of the software is supervised by expert statisticians.

The body of knowledge of statistical science applied to medical research is conventionally presented on a background of basic principles illustrated with worked examples (Armitage and Berry, 1994; Altman, 1991; Bland, 1996). Commonly used statistical software packages, however, routinely present information to support the data preparation aspects of their functions, and provide more statistically oriented materials as separate tutorials with associated data files (SPSS Corporation, 2000; Minitab Corporation 1999; Stata Corporation, 1999). The first line of knowledge support offered in most software is the item in the help system that is associated with a menu item or context-specific position in a document. None of SPSS, Minitab or Stata provides context-sensitive help in reports, but SPSS and Minitab provide help linked with menu items and dialog boxes that interact with the user over statistical operations. Context sensitive statistical knowledge support was developed for the software for this thesis; it was applied to menu items, dialog boxes and reports.

In order to aid statistical understanding, worked examples with inferences were incorporated into the software for this thesis. Such inferences were usually made to conclude with a confidence interval around the calculated effect of the variable(s) observed; thus encouraging inference to be made from confidence intervals over and above calculated probabilities. Encouragement of inference from confidence intervals is in line with contemporary peer-reviewed commentary on priorities for improving statistical practice in medical research (Altman, 1994).

For the reasons described above, the software of this thesis, unlike traditional statistical software, puts statistical knowledge support close to the steps by which a researcher might select and apply a statistical function.

As statistics evolves, certain of its methods sit more comfortably than others in what medical statisticians might agree is the essential statistical armamentarium for medical research. General statistical textbooks applied to medical research vary in the level of detail that they provide, but they agree closely on the overview of methods that they specify or imply to be core (Armitage and Berry, 1994; Altman, 1991; Bland, 1996; Colton, 1974; Petrie, 1990; Campbell and Machin, 1999). Bland (1996) specifies, in the form of decision matrices, an overview of statistical methods appropriate for common statistical goals in medical research. Other authors, for example Altman (1991), use a goal-oriented structure, thereby implying that the statistical methods they use to address these goals are the methods of first or only choice. If the breadth and depth of general texts in medical statistics are used to define the appropriateness of coverage of statistical methods for medical research, then the software of this thesis is more appropriately focused to medical research than are other statistical software packages. The relatively greater medical focus of the software of this thesis, when compared with SPSS, Minitab and Stata, has been demonstrated here.

Statistical software development that is driven by commercial market forces is likely to become targeted more to industrial than to academic or health audiences. For example, SPSS no longer expand their original acronym to "Statistics Package for the Social Sciences" and the current banner across their web site promotes the use of SPSS software for industrial data mining (SPSS Corporation, 2000). The focus of some commercial statistical software remains stable, independently of commercial drives. For example, Genstat software reflects the research interests of its main authors at Rothamstead Experimental Station in Hertfordshire, England; consequently, it has state-of-the-art regression and analysis of variance functions, but nonparametric methods are scarcely covered (Numerical Algorithms Group Ltd., 2000). The software of this thesis was developed with a clear focus upon the statistical needs of medical researchers (primarily non-statisticians). This focus has remained constant for the past eleven years and the number of users of StatsDirect software grew faster in the third quarter of 2000 than in any previous quarter throughout this work.

The potential of the software of this thesis to improve statistical appreciation and practice has been evaluated formatively (Scriven, 1967, 1991, 1996) over the eleven years of its development up to the time of writing this thesis. Formative observations included coverage and focus of the software's statistical functions, rational appropriateness of the support given to users' statistical knowledge by the software, uptake of the software, ad-hoc feedback from users and external reviews. Evolving, engineered processes, such as the development of the software for this thesis, are better suited to formative than to summative evaluation (Scriven, 1991, 1996). Ironically, this work was engineered to support the measurement of uncertainty within the largely summative evaluation methods that underpin medical research. Useful summative evaluation of the effects of the software of this thesis upon statistical appreciation and practice was not feasible within the resources of the project. Even given the resources to make detailed relevant psychometric studies, there would most likely be problems in recruiting a large enough sample of subjects to test in a large enough number of research scenarios. The assumptions made in these circumstances would restrict the generalisability of the summative conclusions. For these reasons, the formative approach to evaluation taken in this thesis was appropriate.

Lessons learnt from developing the software for this thesis

"The significant problems we face cannot be solved at the same level of thinking we were at when we created them".
Albert Einstein (1879 - 1955)

Aspects of the development of the software for this thesis that the author would have done differently, given hindsight and the relevant opportunities, are examined in this section.

The software comprises around five million bytes of source code and six million bytes of help system source in rich text format. At the outset of the work, in 1989, a development as large as this was not anticipated and thus the project was not optimally managed to deliver the software presented herein. This situation was partly unavoidable due to essential iterative refinement of the definition of the need for the software in response to feedback from its users. The process could, however, have been usefully expedited with more programming resource. As sole author of the software and supporting materials, the author of this thesis fitted all developments in between other commitments of a medical career. Given the same situation again, the author would still aim to be sole author but would endeavour to do this work full-time for a substantial period.

The development of the software for this thesis reflected the development of the author's knowledge and skills in both computer software engineering and statistics applied to medical research. A key strength of this co-development has been the integration, in one mind, of an overview of the problem and the engineering. This overview provided a constructive insight into the statistical appreciation and practice of medical researchers. A weakness of the single-author approach is that development priorities can easily become unbalanced. For example, with the initial developments of the software for this thesis (Arcus Pro-Stat for DOS), the author completely wrote a spreadsheet and help system (viewer and compiler). Both the help system and the spreadsheet were more efficient than third party alternatives at the time, but the author’s time could have been more usefully spent

on computational statistical work, waiting for third party help system and spreadsheet components to evolve. In later software developments for Microsoft Windows, the author used third party components when and where it was more efficient to do so. In the software (Arcus QuickStat) later developed for the 16-bit Windows platform, both the spreadsheet and the report editor were third party components. For the 32-bit Windows application presented as the software for this thesis (StatsDirect), it was necessary to write a report object, as third party alternatives performed poorly. StatsDirect employs third party spreadsheet (Tidestone Formula One) and charting (Tidestone First Impression) components (Tidestone Corporation, 1999).

Computational statistics is a speciality of computer software engineering and of statistics. Since the 1960s, statistical computing algorithms have been published and debated in scientific literature, mainly in the FORTRAN language. The author chose to keep parity with this syntax by using a combination of BASIC and FORTRAN languages for the computational statistical work in this thesis. BASIC and FORTRAN are similar in their logical structures and level of verbosity. The author found the resulting code easy to read in terms of the flow of calculations, and therefore initially used few comments in the code. For each function, the author kept paper notes with copies of relevant literature. Many of the algorithms were revised later to overcome precision deficits or instabilities caused by reliance upon published algorithms. At this point, it would have been helpful for references to literature, together with solutions to any errors in that literature, to be recorded in comments in the source code. This was particularly apparent when the precision of all floating-point calculations in the software was increased from 32-bit to 64-bit. The author eventually used the source code as the principal location for recording notes on relevant computational statistical work.

In the early stages of this work, the author placed too much confidence in the ability of the peer review process to eradicate errors in published algorithms and formulae. The frequency of publication of "Remarks" (usually corrections) to algorithms in the former algorithms section of Appendix C of The Journal of The

Royal Statistical Society exemplifies this problem (Young and Minder 1974; Thomas, 1979; Chou, 1985; Boys, 1989; Geodhart and Jansen, 1992). Errors in textbooks, for example formula 2.28 on page 69 of Hosmer and Lemeshow (1999) should have a numerator that is the square of the terms shown without the final c, are slower to be corrected by published errata unless the textbook is well supported via Internet. Numerical algorithms for the software for this thesis were latterly tested around expected results calculated using as many different sources as possible, preferably involving mathematical statisticians working in the relevant fields. Simulation testing with randomly generated data was also employed to look for errors such as non-convergence or results out with known boundaries.

The software of this thesis was initially distributed as shareware (using magnetic media); an arrangement whereby users could choose to receive the latest version of the software plus a printed manual by registering. For the next phase of development, in order to remove the distribution layer, a publisher was engaged. The author had not anticipated that involvement of a publisher would make interaction with users of the software more difficult and remove the ability to directly offer low or no cost distribution in relevant circumstances. The relevant publishing agreement covers only 16-bit Windows platforms, and therefore does not relate to the 32-bit software presented with this thesis. By 1999, adequate access to Internet email and the World Wide Web was almost universal amongst medical researchers and 32-bit Windows platforms had become the most common operating systems for personal computers. In the third quarter of 1999, StatsDirect software for Windows 98, NT and 2000 was released directly to end-users via www.statsdirect.com. Through this work, the author has learned that progress in the development of statistical software is best realised through as close a connection as possible between software engineers, computational statisticians and researchers applying statistical methods.

Feedback from the users of the software of this thesis has continuously fed improvement in the work, often through much iteration. This feedback revealed common errors in statistical appreciation, and the software was adapted to help its users to avoid such errors. For example, after it became apparent that the two by two chi-square function was being used to answer questions more appropriately addressed by inference from a confidence interval for either odds ratio or relative risk, the two by two chi-square function was adapted to prompt the user to specify the type of study that gave rise to their data. The two by two chi-square function in the software of this thesis was eventually written to provide risk statistics relevant to the type of study specified. It is likely that further benefits would have been gained if more had been done to capture constructive feedback such as this.

Plans for further research and development

"Real knowledge is to know the extent of one's ignorance."
Confucius (551 - 479 BC)

The work of this thesis has highlighted statistical goals for which methods are not well developed. An example of this is analysis of agreement for categorical data, in which Scott's pi coefficient has a number of properties that make it more suitable than Cohen's kappa for general inference about agreement. Methods for constructing confidence intervals for Scott's pi are not well studied for the case of more than two observed categories (Zwick, 1988; Donner and Eliasziw 1992). Bootstrap methods may be appropriate here, but careful study of many simulations is required before implementation in the software can be justified (for routine use) (Davidson and Hinkley, 1999). The author plans to extend the work of this thesis to support primary statistical research in areas of medical research need, such as the analysis of agreement.

The coverage of statistical methods in the software of this thesis is due to be extended as development priorities are identified and resources permit. Current development plans include methods for the comparison of ROC curves, cluster analyses, conditional logistic regression for case-control analysis of matched data and generalized categorical regression functions including Poisson regression for multivariate modelling of relative risk from cohort studies. The recent integration of International Mathematical and Statistical Libraries for unrestricted use in software written using the Compaq Visual FORTRAN compiler may speed up these developments (Compaq Corporation, 2000). Such libraries of numerical algorithms have previously been unavailable for use in the work of this thesis.

The software of this thesis is used in educational settings where statisticians facilitate statistical learning. Based upon the findings presented here, it is likely that scientifically constructive information would be gained by structured, systematic feedback on the use of statistical software in such environments. The author plans to actively gather more feedback from statistical educators and

students who use the software of this thesis as a learning tool. In this way, the work of this thesis may detect and respond to indicators of statistical misconception. Another planned educational development is the publication, at the www.statsdirect.com web site, of print-ready introductory materials for students who are new to statistics and want to use the software of this thesis in their learning process.

Conclusions

"When I am working on a problem I never think of beauty, I only think about how to solve the problem. But when I have finished, if the solution is not beautiful, I know it is wrong".
R Buckminster Fuller (1895 -1983)

The work of this thesis has given rise to a statistical computing resource for medical research. This concluding section presents the original contribution and position of this work in the areas where statistical, computing and medical knowledge needs overlap.

Most of the statistical methods covered here are based upon probabilistic sampling and model-based testing. Such methods have profoundly influenced scientific practice (Stigler, 1986). Medicine has been both a driver and a slow adopter of statistical methods, and this paradox has been widely reported for the past three decades (Cohrane, 1972; Stigler, 1986; Altman, 1991). The persistence of poor statistical practice in medical research has been attributed variously to educational shortfalls, perverse incentives for publication, and the misuse of computer software (Altman, 1994). An original approach to this problem, taken in the work of this thesis, was to develop statistical computer software specifically for non-statistician medical researchers. In order to achieve this, meeting the statistical knowledge needs of medical researchers was treated as a routine function of the software, acknowledging that medical researchers may need to re-learn statistical principles, and not rely upon previous learning. The statistical knowledge-support function of the software produced in the work of this thesis was found to be unusually comprehensive; as evidenced by the results presented here and by peer review (Fitzpatrick, 2000; Freemantle, 1998a, 2000; Elliott, 1998; Sithers, 1997).

The growth in networked computing environments, particularly from the late 1980s onward, has given rise to new types of data that currently occupy much original statistical research (Wegman, 2000). Consequently, traditional medical research data are becoming less prominent than other data as stimuli for research

and development in mathematical statistics. The new huge data sets, from increasingly ubiquitous and powerful computer systems, are slowly emerging in healthcare environments, but samples that are small to medium size in statistical terms are likely to remain essential to medical research. This essential constraint on sample sizes is partly due to ethics. For example, most people would agree that it is unethical to study an intervention of unknown efficacy and/or safety in more human subjects than are necessary to accept or refute a hypothesis about the efficacy and/or safety of the intervention (Beauchamp and Childress, 1994). As the ethical basis for medical research is unlikely to change, statistical methods for small to medium sized samples are likely to retain their key role in medical research for the foreseeable future. The work of this thesis has provided a self-sustaining statistical computing resource that focuses upon the small to medium size sample methods that are core to medical research. This resource has successfully been a platform for the dissemination of new and improved statistical methods for medical research (e.g. Newcombe, 1998a).

Irrespective of developments in statistical computing, the medical research community continues to be criticised for failing to appropriately use well-established statistical methods (Altman, 1994). This implies a failure of statistical appreciation and practice by medical researchers. Software designed to support statistical practice should therefore support not only statistical computation, but also statistical knowledge. Principles of knowledge management have arisen formally in the fields of management and organisational theory (Davenport and Prusak, 1998). The same principles are shaping the development of computer systems that will give rise to huge data sets: a subject currently absorbing much research and development activity in mathematical statistics and computing (Microsoft Research, 2000; Wegman, 2000). Two important dangers are presented by such developments. The first danger is that computational tools can be misconceived as a true representation of statistical science, and thereby distract attention from more important statistical goals. A symptom of this problem is already apparent in the structure of some introductory statistical texts, such as Streiner and Norman (1979), which reflect the facilities of common computational

tools more than the categories of common statistical goals and tasks. The same criticism could be levelled at the "Analysis" menu structure in the software of this thesis; there is room for improvement. The second danger is neglect of the statistical knowledge needs of non-statistician investigators. For the reasons stated in the introduction chapter, mathematical statistics may become more removed from the statistical problems of medical researchers, a situation that would increase the knowledge gap associated with poor statistical practice in medical research. The work of this thesis has produced an original framework to support both statistical knowledge and calculation in routine medical research, and this has become widely used.

The author began the work of this thesis by studying statistics from a combination of applied and computational perspectives. From this experience, computer software was engineered to support the appropriate use of statistical methods in medical research. Feedback from users of this software has since informed the continuous development of methods to support statistical knowledge. From the results presented here, the author concludes that the computing resource produced in the work of this thesis has substantial potential to improve statistical appreciation and practice in medical research.

References

Aalen OO. Non parametric inference for a family of counting processes. *Annals of Statistics* 1978;6:701-726.

Abramowitz M, Stegun IA. Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables (9th printing with corrections). New York: Wiley 1972.

Adler E. Review of Arcus ProStat. *Personal Computer World* (page 306). VNU Business Publications March 1994.

Agresti A. An introduction to categorical data analysis. New York: Wiley 1996.

Altman DG. Practical Statistics for Medical Research. Chapman and Hall 1991.

Altman DG. The scandal of poor medical research. *British Medical Journal* 1994;308:283-4.

Anbar D. On estimating the difference between two probabilities, with special reference to clinical trials. *Biometrics* 1983;39:257-262.

Andersen PK et al.. *Statistical models based on counting processes*. New York: Springer-Verlag 1993.

Armitage P, Berry G. *Statistical Methods in Medical Research* (3rd edition). Blackwell 1994.

Aziz N, Buchan I, *et al*. Sperm deformity index: a reliable predictor of the outcome of fertilization in vitro. *Fertility and Sterility* 1996;66(6):1000-8.

Bailey NTJ. Mathematics, Statistics and Systems for Health. New York: Wiley 1977.

Beauchamp TL, Childress JF. *Principles of Biomedical Ethics.* New York; Oxford: Oxford University Press 1994.

Berger RL. Remark R86 on algorithm AS152, Cumulative hypergeometric probability. *Applied Statistics* 1991;40(2):374-5.

Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics*. Wiley 1980.

Berkson J, Gage RP. Calculation of survival rates for cancer. *Proceedings of Staff Meetings of the Mayo Clinic* 1950;25:270-286.

Berry KJ, Mielke PW, Cran GW. Remark R83 further to AS64. *Applied Statistics* 1990;39(2).

Best DJ, Gipps PG. AS71, Upper Tail Probabilities of Kendall's Tau. *Applied Statistics* 1974;23(1).

Best DJ, Roberts DE. AS89, Upper Tail Probabilities of Spearman's Rho. *Applied Statistics* 1975;24(3).

Best DJ, Roberts DE. AS91, The Percentage Points of the Chi² Distribution. *Applied Statistics* 1975;24(3).

Bland M, Altman DG. Statistical Methods for Assessing the Difference Between Two Methods of Measurement. *Lancet* 1986;i:307-310.

Bland M. *An Introduction to Medical Statistics* (2nd edition). Oxford Medical Publications 1996.

Bland MJ, Altman DG. Measurement Error and correlation coefficients. *British Medical Journal* 1996;313:41-2.

Bland MJ, Altman DG. Statistics Notes: Measurement error. *British Medical Journal* 1996;312:1654.

Bland MJ, Altman DG. Statistics Notes: Transforming data. *British Medical Journal* 1996;312:770.

Boys R. Remark R80 on AS76, an integral useful in calculating non-central t and bivariate normal probabilities. *Applied Statistics* 1989;38(3):580-2.

Breslow NE. Covariance analysis of censored survival data. *Biometrics* 1974;30:89-99.

Breslow NE, Day NE. Statistical Methods in Cancer Research: Vol. I - The Analysis of Case Control Studies. Lyon: International Agency for Research on Cancer 1980.

Breslow NE, Day NE. Statistical Methods in Cancer Research: Vol. II - The Design and Analysis of Cohort Studies. Lyon: International Agency for Research on Cancer 1987.

Brookmeyer R, Crowley J. A confidence interval for the median survival time. *Biometrics* 1982;38:29-41.

Brown MB, Forsythe AB. Robust tests for the equality of variances. *Journal of the American Statistical Association* 1974;69:364-7.

Bryson MC, Johnson ME. The incidence of monotone likelihood in the Cox model. *Technometrics* 1981;23:381-4.

Bundred PE, Buchan IE, Kichiner DJ. A population-based study demonstrating an increase in the number of overweight and obese children between 1989 and 1998. Archives of Disease in Children 2000;82(Supplement 1):A35-36.

Bundred PE, Kitchener DJ, Buchan IE. An increase in the number of overweight and obese children between 1989 and 1998: A population based series of cross sectional studies. British Medical Journal 2000 in press.

Campbell MJ, Machin D. *Medical Statistics: a commonsense approach* (3rd edition). Chichester: Wiley 1999.

Casagrande JT, Pike MC, Smith PG. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics* 1978;34:483-486.

Chalmers I, Dickersin K, Chalmers TC. Getting to grips with Archie Cochrane's agenda. *British Medical Journal* 1992;305:786-88.

Chan TF. Algorithm 581. Singular value decomposition of a general rectangular matrix. *Transactions on Mathematical Software* 1982;8(1):84-88.

Chou YM. Remark R55 on AS76, an integral useful in calculating non-central t and bivariate normal probabilities. *Applied Statistics* 1985;34(1):100-1.

Cochran W, Cox G. *Experimental Designs* (2nd edition). Wiley 1957.

Cody WJ, Hillstrom KE. Chebyshev Approximations for the Natural Logarithm of the Gamma Function. *Mathematics of Computation* 1967;21:198-203.

Coe PR, Tamhane AC. Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. *Communications in Statistics and Simulation* 1993;22(4):925-938.

Collett D. *Modelling Survival Data in Medical Research*. Chapman and Hall 1994.

Colton T. *Statistics in Medicine*. Little Brown & Co 1974.

Cochrane AL. *Effectiveness and Efficiency. Random Reflections on Health Services*. London: Nuffield Provincial Hospitals Trust 1972.

Compaq Corporation. *Compaq Visual FORTRAN* (version 1.6). 2000: www.digital.com/fortran.

Conover WJ, *Practical Nonparametric Statistics* (3rd edition). Wiley 1999.

Copenhaver MD, Holland BS. Computation of the distribution of the maximum Studentized range statistic with application to multiple significance testing of simple effects. *Journal of Statistical Computing and Simulation* 1988;30:1-15.

Cox DR. Regression models and life tables. *Journal of the Royal Statistical Society* 1972;B34:187-220.

Cox DR, Oakes D. *Analysis of survival data*. London: Chapman and Hall 1984.

Cox DR, Snell EJ. *The Analysis of Binary Data* (2nd edition). Chapman and Hall 1989.

Cran GW, Martin KJ. Thomas GE. R19 and AS 109 further to AS 63 and AS 64. *Applied Statistics* 1977;26(1).

Critchlow DE, Fligner MA. On distribution-free multiple comparisons in the one-way analysis of variance. *Communications of Statistical Theory and Methods* 1991;20:127-139.

Cronbach L, Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 1951;16(3):297-333.

Cuzick J. A Wilcoxon-Type Test for Trend. *Statistics in Medicine* 1985;4:87-89.

Davenport TH and Prusak L. *Working Knowledge: How Organizations Manage What They Know*. Boston: Harvard Business School Press 1998.

David HA. *Order Statistics* (2nd edition). New York: John Wiley & Sons 1981.

Davidson AC, Hinkley DV. *Bootstrap Methods and their Application*. Cambridge University Press 1999.

DerSimonian R, Laird N. Meta-analysis in Clinical Trials. *Controlled Clinical Trials* 1986;7:177-188.

Dinneen LC, Blakesley BC. AS62, A Generator for the Sampling Distribution of the Mann-Whitney U Statistic. *Applied Statistics* 1973;22(2).

Donner A, Eliasziw M. A goodness of fit approach to inference procedures for the Kappa statistic: CI construction, significance testing and sample size estimation. *Statistics in Medicine* 1992;11:511-519.

Dorsey EN. The velocity of light. Transactions of the American Philosophical Society 1944;34:1-110.

Dowrick C, Buchan I. Twelve month outcome of depression in general practice: does detection or disclosure make a difference? *British Medical Journal* 1995;311(7015):1274-6.

Draper NR, Smith H. *Applied Regression Analysis* (3rd edition). New York: Wiley 1998.

Dupont WD. Power calculations for matched case-control studies. *Biometrics* 1988;44:1157-1168.

Dupont WD. Power and Sample size calculations. *Controlled Clinical Trials* 1990;11:116-128.

Egger M, *et al*. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 1997;315:629-634.

Elliott C. Review: Arcus QuickStat Biomedical. *Life Sciences Educational Computing* 1998;9(2):24-25.

Ernst M. *The Calculating machines (Die Rechenmaschinen): their history and development.* Cambridge Massachusetts: MIT Press 1992.

Everitt B, Dunn G. *Applied Multivariate Data Analysis*. Edward Arnold 1991.

Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *Journal of the American Medical Association* 1992;268:2420-25.

Eyler JM. *Victorian Social Medicine: the ideas and methods of William Farr*. Baltimore: Johns Hopkins University Press, 1979.

Finney DJ. *Probit Analysis*. Cambridge University Press 1971.

Finney DJ. *Statistical Method in Biological Assay*. Charles Griffin & Co. 1978.

Fitzpatrick V. Software Review: StatsDirect 1.615. *HMS Beagle (BioMedNet)* 2000 (23 June);81:news.bmn.com/hmsbeagle/81/reviews/sreview.

Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: A critique. *Journal of Clinical Epidemiology* 1991;44:127-39.

Fleiss JL. Confidence intervals for the odds ratio in case-control studies: the state of the art. *Journal of Chronic Diseases* 1979;32:69-77.

Fleiss JL. Statistical Methods for Rates and Proportions (2nd edition). Wiley 1981.

Fleiss JL. The statistical basis of meta-analysis. *Statistical Methods in Medical Research* 1993;2:121-145.

Fog A. *Pseudo random number generators*. 2000: www.agner.org/random.

Freeman P. Review of Arcus Pro-Stat. *CTI Centre for Medicine Update* April 1995.

Freeman P. Review of Arcus Pro-Stat. *Journal for Audiovisual Media in Medicine* 1995;18(2).

Freemantle N. Personal communication of data from a meta-analysis 1998.

Freemantle N. Review of Arcus QuickStat. *British Medical Journal* 1998;316:159.

Freemantle N. Review of StatsDirect. *British Medical Journal* 2000 in press.

Gardner MJ, Altman DG. Statistics with Confidence - Confidence Intervals and Statistical Guidelines. British Medical Journal 1989.

Gart JJ, Nam J. Approximate interval estimation for the difference in binomial parameters: correction for skewness and extension to multiple tables. *Biometrics* 1990;46:637-643.

Gart JJ, Nam J. Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness. *Biometrics* 1988;44:323-338.

Gentleman WM. Basic procedures for large, sparse or weighted linear least squares problems. *Applied Statistics* 1974;23:448-454.

Geodhart PW, Jansen MJW. Remark R89 on AS76, an integral useful in calculating non-central t and bivariate normal probabilities. *Applied Statistics* 1992;41(2):496-7.

Gleason JR. An accurate, non-iterative approximation for Studentized range quantiles. *Computational Statistics and Data Analysis* 1999;31(2):147-158.

Golub GH, Van Loan CF. *Matrix Computations*. Baltimore, Maryland: Johns Hopkins University Press 1983.

Goodman LA, Kruskal WH. Measures of association for cross-classifications III: Approximate sampling theory. *Journal of the American Statistical Association* 1963;58:310-364.

Greenland S, Robins JM. Estimation of common effect parameter from sparse follow up data. *Biometrics* 1985;41:55-68.

Greenland S. RE: A simple method to calculate the confidence interval of a standardized mortality ratio. *American Journal of Epidemiology* 1990;133(2):212-213.

Greenland S, Salvan A. Bias in the one-step method for pooling study results. *Statistics in Medicine* 1990;9:247-252.

Greenwood M. The natural duration of cancer. *Reports on Public Health and Medical Subjects.* London: Her Majesty's Stationery Office 1926;33:1-26.

Greenwood M. *The medical dictator and other biographical studies* (reprint of 1936 with an introduction by Austin Bradford Hill). London: The Keyenes Press 1986.

Hanka R. Personal communication 1997.

Hanley JA, McNeil BJ. The meaning and use of area under a Receiver Operating Characteristic (ROC) curve. *Radiology* 1982;143:29-36.

Harding EF. An Efficient Minimal Storage Procedure for Calculating the Mann-Whitney U, Generalised U and Similar Distributions. *Applied Statistics* 1983;33.

Harris EK, Albert A. *Survivorship analysis for clinical studies*. New York: Dekker 1991.

Harter HL. Expected values of normal order statistics. *Biometrika* 1961;48:151-165.

Haynes RB, Sackett DL. personal communication. McMaster University 1993.

Hedges, Olkin. *Statistical methods for meta-analysis*. London: Academic Press 1985.

Hesse M. Doomsday land-measures in Suffolk. *Landscape History*. in press 2000.

Hill AB. *A Short Textbook of Medical Statistics* (11th edition). London: Hodder and Stoughton 1984.

Hill AB, Hill ID. *Bradford Hill's Principles of Medical Statistics.* London: Edward Arnold 1991.

Hill AB. *Principles of medical statistics (The Lancet Postgraduate Series).* London: The Lancet 1937.

Hill GW. Student's t-Quantiles (Algorithm 396). *Communications of the Association for Computing Machinery* 1970;13:619-620.

Hill ID. AS66, The Normal Integral. *Applied Statistics* 1973;22(3).

Hocking DC. *The Analysis of Linear Models*. Monterrey, California: Brookes-Cole 1985.

Hogg RV, Tanis EA. *Probability and Statistical Inference* (4th edition). New York: MacMillan 1993.

Hollander M, Wolfe DA. *Non-parametric Statistical Methods* (2nd edition). New York: Wiley 1999.

Honeyball J. Real World Computing: 32-Bit Windows (Arcus). *PC Pro Magazine* August 1997:294-295.

Hosmer DW, Lemeshow S. *Applied Logistic Regression*. New York: Wiley 1989.

Hosmer DW, Lemeshow S. *Applied Survival Analysis*. New York: Wiley 1999.

Hotelling H, Pabst MR. Rank correlation and tests of significance involving no assumption of normality. *The Annals of Mathematical Statistics* 1936;7:29-43.

Hsu JC. *Multiple Comparisons*. Chapman and Hall 1996.

IEEE Standard for Binary Floating Point Numbers, ANSI/IEEE Std 754. New York: Institute of Electrical and Electronics Engineers (IEEE) 1985.

Ioannidis JP, *et al*. Early or deferred Zidovudine therapy in HIV-infected patients without and AIDS-defining illness. *Annals of Internal Medicine* 1995;122:856-66.

Iman RL, Davenport JM. New approximations to the exact distribution of the Kruskal-Wallis test statistic. *Communications in Statistics - Theory and Methods* 1976;A5:1335-1348.

Iman RL, Davenport JM. Approximations to the critical region of the Friedman statistic. *Communications in Statistics - Theory and Methods* 1980;A9:571-595.

Jenkins J. Software Tools for Research. *The Doctors' Post* 1995;88:7.

Johnson NL, Kotz S. *Discrete Distributions*. Boston: Houghton Mifflin Company 1969.

Johnson NL, Kotz S. Continuous Univariate Distributions (1 and 2). New York: Wiley 1970.

Johnson RA, Wichern DW. *Applied Multivariate Statistical Methods* (4th edition). London: Prentice-Hall 1998.

Kalbfleisch JD, Prentice RL. *Statistical Analysis of Failure Time Data*. New York: Wiley 1980.

Kalbfleisch JD, Prentice RL. Marginal likelihoods based on Cox's regression and life model. *Biometrika* 1973;60(2):267-278.

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958;53:457-481.

Kendall MG, Gibbons JD. *Rank Correlation Methods* (5th edition). London: Arnold 1990.

Keuls M. The use of 'Studentized range' in connection with analysis of variance. *Euphytica* 1952;1:112-122.

Kim PJ, Jennrich RI. Tables of the exact sampling distribution of the two sample Kolmogorov-Smirnov criterion. in *Selected Tables in Mathematical Statistics* (Vol 1). Providence: American Mathematical Society 1973.

Kleinbaum DG, et al. Applied Regression Analysis and Other Multivariable Methods (3rd edition). Duxbury Press 1998.

Knusel L. Computation of the Chi-square and Poisson distribution. *SIAM Journal on Scientific and Statistical Computing* 1986;7:1022-1036.

Knuth DE. *The Art of Computer Programming : Seminumerical Algorithms* (Art of Computer Programming, Vol 2, 2nd edition). Reading, Massachusetts: Addison Wesley 1997.

Knuth DE. *The Art of Computer Programming : Sorting and Searching* (Art of Computer Programming, Vol 3, 2nd edition). Reading, Massachusetts: Addison Wesley 1998.

Koehler KJ, Larnz K. An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association* 1980;75:336-344.

Krzanowski WJ. *Principles of Multivariate Analysis*. Oxford: Oxford University Press 1988.

Landis R, Koch G. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.

Last JM. *A Dictionary of Epidemiology*. New York: Oxford University Press 1995.

Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine* 1988;318(26):1728-33.

Lawless JF. Statistical Models and Methods for Lifetime Data. New York: Wiley 1982.

Le CT. *Applied survival analysis*. New York: John Wiley and Sons 1997.

Lee N. ABC of Medical Computing. *British Medical Journal* 1995;311:617.

Lenth RV. AS243, Cumulative distribution function of the non-central t-distribution. *Applied Statistics* 1989;38(1).

Leung HM, Kupper LL. Comparison of confidence intervals for attributable risk. *Biometrics* 1981;37:293-302.

Lewis T. *Research in medicine and other addresses* (2nd edition). London: Lewis & Company 1945.

Liddell FDK. Simplified exact analysis of case-referent studies; matched pairs; dichotomous exposure. *Journal of Epidemiology and Community Health* 1983;37:82-84.

Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. *British Medical Journal* 1996;313:603-7.

Longley JW. An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical Association* 1967;62:819-841.

Lordan C. Reviews: Arcus ProStat. *Life Sciences Educational Computing* 1995;6(2):16-17.

Lund RE, Lund JR. AS190, Probabilities and Upper Quantiles for the Studentized Range. *Applied Statistics* 1983;34.

MacTutor History of Mathematics Archive. St Andrews University 2000: www.groups.dcs.st-and.ac.uk/~history.

Macleod AJ. AS245, A Robust and Reliable Algorithm for the Logarithm of the Gamma Function. *Applied Statistics* 1989;38(2).

Majumder KL, Bhattcharjee GP. AS63, The Incomplete Beta Integral. *Applied Statistics* 1973;22(3).

Majumder KL, Bhattcharjee GP. AS64, Inverse of the Incomplete Beta Function Ratio. *Applied Statistics* 1973;22(3).

Makuch RW, Escobar M.. AS262, A Two Sample Test for Incomplete Multivariate Data. *Applied Statistics* 1991;40(1).

Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies. *Journal of the National Cancer Institute* 1959;22:719-748.

Marsaglia G. Monkey tests for random number generators. *Computers and Mathematics with Applications* 1993;26:1-10.

Marsaglia G. *DIEHARD: A battery of tests of randomness*. 1997: stat.fsu.edu/pub/diehard.

Maxwell A E, Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry* 1970;116:651-5.

McClave JT, Deitrich FH. *Statistics* (5th edition). Macmillan 1991.

McCullough BD, Wilson B. On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics and Data Analysis* 1999;31:27-37.

McCullagh P, Nelder JA. *Generalised Linear Models* (2nd edition). Chapman and Hall 1989.

McDowell I, Newell C. Measuring Health: a guide to rating scales and questionnaires. Oxford University Press 1987.

McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1996;1(1):30-46.

McManus C. Engineering quality in health care. *Quality in Health Care* 1996;5:127.

Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal* 1948;ii:769-82.

Mee RW. Confidence bounds for the difference between two probabilities. *Biometrics* 1984;40:1175-1176.

Mehta CR, Patel NR. A network algorithm for performing Fisher's exact test in r c contingency tables. *Journal of the American Statistical Association* 1983;78:427-34.

Mehta CR, Patel NR. Algorithm 643: FEXACT: A FORTRAN subroutine for Fisher's exact test on unordered r x c contingency tables. *ACM Transactions on Mathematical Software* 1986;12:154-61.

Mehta CR, Patel NR. A hybrid algorithm for Fisher's exact test in unordered r x c contingency tables. *Communications in Statistics, Series A* 1986;15:387-404.

Meinert CL. *Clinical Trials: Design, Conduct and Analysis.* New York: Oxford University Press 1986.

Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 1978;8:283-98.

Microdexterity Corporation. *Microdexterity Stamina ASM Library.* 1999: www.microdexterity.com.

Microsoft Corporation. *Microsoft Visual Studio.* 1998: www.microsoft.com.

Microsoft Research. *Data Management, Exploration and Mining.* 2000: www.research.microsoft.com/dmx.

Miettinen OS, Nurminen M. Comparative analysis of two rates. *Statistics in Medicine* 1985;4:213-226.

Miller JR. *Survival Analysis*. New York: Wiley 1981.

Miller RG (jnr.). *Simultaneous Statistical Inference* (2nd edition). Springer-Verlag 1981.

Minitab Corporation. *Minitab for Windows* (release 12.23). 1999: www.minitab.com.

Mitchell M. Review of StatsDirect. *Journal of the American Medical Association* 2000;284(15), 1988.

Morris AH. Algorithm 708, Incomplete Beta Function. *Transactions on Mathematical Software* 1992;18(3):360-373.

Mulrow CD, Oxman AD (eds.). *Cochrane Collaboration Handbook*. Oxford: Cochrane Collaboration 1996.

Nelson W. Theory and applications of hazard plotting for censored failure data. *Technometrics* 1972;14:945-966.

Neumann N. Some Procedures for Calculating the Distributions of Elementary Non-parametric Test Statistics. *Statistical Software Newsletter* 1988;14(3).

Newcombe R. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* 1998;17:2635-2650.

Newcombe R. Interval estimation for the difference between independent proportions. *Statistics in Medicine* 1998;17:873-890.

Newcombe R. Two sided confidence intervals for the single proportion: a comparative evaluation of seven methods. *Statistics in Medicine* 1998;17:857-872.

Newman D. The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika* 1939;31:20-30.

Nikiforov AM. AS288, Exact Smirnov two sample tests for arbitrary distributions. *Applied Statistics* 1994;43(1):265-284.

Norman GR, Streiner DL. *PDQ Statistics* (2nd edition). St. Louis: Mosby-Year Book 1997.

Numerical Algorithms Group Ltd.. *Genstat* (version 5, release 4.1). 2000: www.nag.co.uk.

Odeh RE, Evans JO. AS70, Percentage Points of the Normal Distribution. *Applied Statistics* 1974;23.

Onion CW, Dutton CE, Walley T, Turnbull CJ, Dunne WT, Buchan IE. Local clinical guidelines: description and evaluation of a participative method for their development and implementation. *Family Practice* 1996;13(1):28-34.

Owen DB. A special case of the bivariate non-central t-distribution. *Biometrika* 1965;52:437-446.

Pearson & Hartley. *Biometrika tables for statisticians* (Volumes I and II, 3rd edition). Cambridge University Press 1970.

Peterson AV Jnr.. Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *Journal of the American Statistical Association* 1977;72:854-858.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Part I: Introduction and design. *British Journal of Cancer* 1976;34:585-612.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Part II: Analysis and Examples. *British Journal of Cancer* 1977;35:1-39.

Peto R, Peto J. Asymptotically efficient rank invariant procedures. *Journal of the Royal Statistical Society* 1972;A135:185-207.

Petrie A. *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications 1990.

Pike MC, Hill ID. Algorithm 291, Logarithm of the Gamma Function. *Communications of the Association for Computing Machinery* 1966;9:684.

Pirmohamed M et al.. Lack association between schizophrenia and the CYP2D6 gene polymorphisms. *American Journal of Medical Genetics* 1996;67(2):236-7.

Pregibon D. Logistic Regression Diagnostics. *Annals of Statistics* 1981;9:705-724.

Press WH, *et al*. *Numerical Recipes, The Art of Scientific Computing* (2nd edition). Cambridge University Press 1992.

Robins J, Breslow N, Greenland S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large strata models. *Biometrics* 1986;42:311-323.

Rice JA. *Mathematical Statistics and Data Analysis* (2nd edition). Belmont, California: Duxbury Press 1995.

Ross JG. *NonLinear Estimation*. Springer-Verlag New York 1990.

Rothman KJ, Monson RR. Survival in trigeminal neuralgia. *Journal of Chronic Diseases* 1973;26:303-9.

Rothman KJ, Greenland S. *Modern Epidemiology* (2nd edition). Philadelphia: Lippincott-Raven 1998.

Royston JP. AS177, Expected Normal Order Statistics (Exact and Approximate). *Applied Statistics* 1982;31(2):161-165.

Royston JP. AS181, The W Test for Normality. *Applied Statistics* 1982;31(2):176-180.

Royston JP. Remark R94, Shapiro-Wilk normality test and P-value. *Applied Statistics* 1995;44(4).

Royston JP. Remark R69 on AS190, Probabilities and upper quantiles for the Studentized range. *Applied Statistics* 1987.

Rowntree D. *Statistics without tears*. London: Penguin 1991.

Sachdev HPS. Multimedia Review: Arcus QuickStat. *Indian Paediatrics* 1999;36:1075-6.

Sackett DL, *et al*. Interpretation of diagnostic data (5). *Canadian Medical Association Journal* 1983;129:947-975.

Sackett DL, et al. Clinical Epidemiology: A Basic Science for Clinical Medicine. Boston: Little, Brown & Co. 1991.

Sackett DL. On some clinically useful measures of the effects of treatment. *Evidence-Based Medicine* 1996 Jan-Feb;1:37.

Sahai H, Kurshid A. Statistics in epidemiology: methods techniques and applications. CRC Press 1996.

Samra B, Randles RH. A test for correlation based on Kendall's tau. *Communications in Statistics - Theory and Methods* 1988;17:3191-3205.

Santer TJ, Snell MK. Small-sample confidence intervals for p_1-p_2 and p_1/p_2 in 2 x 2 contingency tables. *Journal of the American Statistical Association* 1980;75:386-94.

Sato T. Confidence limits for the common odds ratio based on the asymptotic distribution of the Mantel-Haenszel estimator. *Biometrics* 1990;46:71-80.

Savage MW et al. Vascular reactivity to noradrenaline and neuropeptide Y in the streptozotocin-induced diabetic rat. *European Journal of Clinical Investigation* 1995;25(12):974-9.

Schlesselman JJ. *Case-Control Studies*. New York: Oxford University Press 1982.

Schoenfeld DA, Richter JR. Nomograms for calculating the number of patients needed for a clinical trial with survival as endpoint. *Biometrics* 1982;38:163-70.

Scott WA. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly* 1955;19:321-325.

Scriven MS. *The Methodology of Evaluation* in ed. Stokes RE, *Curriculum Evaluation*, American Educational Research Association Monograph Series on Evaluation No. 1. Chicago: Rand McNally 1967.

Scriven MS. *Evaluation Thesaurus* (3rd edition). Newbury Park: Sage Publications 1991.

Scriven MS. Types of evaluation and types of evaluator. *Evaluation Practice* 1996; 17(2):151-161.

Sellu D. *Practical Personal Computing for Healthcare Professionals* (pages 73 and 423). London: Butterwoth Heineman 1994.

Senn S. Cross-over trials in clinical research. John Wiley 1993.

Shapiro SS, Wilk MB. An analysis of variance test for normality. *Biometrika* 1965;52(3):591-9.

Shea BL. AS239, Chi-square and incomplete gamma integral. *Applied Statistics* 1988;37(3):466-73.

Shea BL. Remark R77 on AS152, Cumulative hypergeometric probability. *Applied Statistics* 1989;38(1):199-204.

Sithers AJ. Review of Arcus QuickStat Biomedical for Windows. *CTI Centre for Medicine Update* 1997.

Snedecor G, Cochran W, Cox D. *Statistical Methods* (8th edition). The Iowa State University Press 1989.

Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Methods in health service research: An introduction to bayesian methods in health technology assessment. *British Medical Journal* 1999;319:508-12.

SPSS Corporation. *SPSS* (version 10). 2000: www.spss.com.

Stampfer MJ, *et al*. A prospective study of postmenopausal hormones and coronary heart disease. *New England Journal of Medicine* 1985;313:1044-49.

Stata Corporation. *Stata* (version 6.0). 1999: www.stata.com.

Statbase. PP98t1: Estimated resident population of the United Kingdom at mid-1998. Office for National Statistics/Government Statistics Service. 2000: www.statistics.gov.uk.

Stigler S, *The History of Statistics, The Measurement of Uncertainty before 1900*, Belknap Press of Harvard University Press, Cambridge Massachusetts and London 1986.

Streiner D, Norman G. Health Measurement Scales: a practical guide to their development and use. *Oxford University Press* 1989.

Stuart A, Ord JK. *Kendall's Advanced Theory of Statistics* (6th edition). London: Edward Arnold 1994.

Tarone RE, Ware J. On distribution-free tests for equality of survival distributions. *Biometrika* 1977;64:165-60.

Thomas DG. AS36, Exact Confidence Limits for the Odds Ratio in a Two by Two Table. *Applied Statistics* 1971;20(1).

Thomas GE. Remark R30 on AS76, An integral useful in calculating non-central t and bivariate normal probabilities. *Applied Statistics* 1979;28(1):113.

Tidestone Corporation. *Tidestone Formula One and First Impression* 1999: www.tidestone.com.

Tukey JW. *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley 1974.

Ulm K. A simple method to calculate the confidence interval of a standardized mortality ratio. *American Journal of Epidemiology* 1990;131(2):373-375.

Vandenbroucke JP. A short note on the history of the randomized controlled trial. *Journal of Chronic Diseases* 1987;40:985-86.

Vandenbroucke JP, Eelkman Rooda HM, Beukers H. Who made John Snow a hero? *American Journal of Epidemiology* 1991;133:967-73.

Vandenbroucke JP. Clinical investigation in the 20th Century: the ascendancy of numerical reasoning. *The Lancet* 1998;352(supplement 2):12-16.

Verrill S, Johnson A. Tables and Large Sample Distribution Theory for Censored Data Correlation Statistics for Testing Normality. *Journal of the American Statistical Association* 1988;83:1192-1197.

Vollset SM. Confidence intervals for a binomial proportion. *Statistics in Medicine* 1993;12:809-824.

Wallenstein S. Some statistical methods useful in circulation research. *Circulation Research* 1980;47(1).

Walter SD. Calculation of attributable risks from epidemiologic data. *International Journal of Epidemiology* 1978;7:175-82.

Wegman EJ. On the eve of the 21st Century: statistical science at a crossroads. *Computational Statistics and Data Analysis* 2000;32:239-243.

Wei LJ, Lachin JM. Two Sample Asymptotically Distribution Free Tests for Incomplete Multivariate Observations. *Journal of the American Statistical Association*. 1984;79:653-661.

Wetherill GB. *Intermediate Statistical Methods*. Chapman Hall 1981.

Wichura MJ. AS 241, The percentage points of the normal distribution. *Applied Statistics* 1988;37:477-483.

Williams G et al. Dissociation of body-temperature and melatonin secretion circadian rhythms in patients with chronic fatigue syndrome. *Clinical Physiology* 1996;16(4):327-337.

Wilson RP, Buchan I, Walley T. Alterations in prescribing by general practitioner fundholders: an observational study. *British Medical Journal* 1995;311(7016):1347-50.

Young JC, Minder CE. AS76, An integral useful in calculating non-central t and bivariate normal probabilities. *Applied Statistics* 1974;23(3):455-457.

Yusuf S, Peto R, Lewis J, Colins R, Sleight P. Beta blockade during and after myocardial infarction. An overview of randomized trials. *Progress in Cardiovascular Disease* 1985;27:335-371.

Zwick R. Another look at inter-rater agreement. *Psychological Bulletin* 1988;103:374-378.

Appendix 1

Instructions for the installation of StatsDirect software either from the StatsDirect web site or from the CD ROM enclosed with this thesis:

From the StatsDirect web site

1. Open the web site www.statsdirect.com/update.htm and enter the details requested.
2. Install the current version of StatsDirect software on a computer running Microsoft Windows 98, NT, 2000 or a later compatible operating system..
3. The first time you run the software, enter your user details as:
   Email/user name: **MDTHESIS**
   Licence key: **please email iain@ukph.org**
4. Run the StatsDirect application; note that the data for the numerical examples presented in this thesis are contained in the "test" workbook.

From the CD-ROM enclosed

1. Take the CD-ROM from the sleeve at the front of this document and place it into a computer running Microsoft Windows 98, NT, 2000 or a later compatible operating system.
2. Run the SetupStatsDirect.EXE application that installs StatsDirect software.
3. The first time you run the software, enter your user details as:
   Email/user name: **MDTHESIS**
   Licence key: **please email iain@ukph.org**
4. Run the StatsDirect application; note that the data for the numerical examples presented in this thesis are contained in the "test" workbook.