WINPEPI PROGRAMS

ETCETERA
MANUAL
(Version 3.26)

Revised Aug 16, 2016

What ETCETERA does

ETCETERA is a WINPEPI program (Abramson 2004, 2011), part of the PEPI suite of computer programs for epidemiologists. ("PEPI" is an acronym for "Programs for EPIdemiologists".)

This program has 35 modules. They perform randomization and random sampling, adjust P values derived from multiple tests, appraise synergism, evaluate scales, compare three or more samples, control unmeasured confounders, and apply procedures concerned with correlation coefficients, large and three-way tables, median and mean polish, simple and multiple linear regression, factorial-design and crossover studies, and Bayes factors.

WORDS OF CAUTION

It is unwise to use a statistical procedure whose use one does not understand. This manual cannot supply this knowledge, and it is certainly no substitute for the basic understanding of statistics and epidemiological thinking that is essential for the wise choice of methods and the correct interpretation of their results.

How to use ETCETERA

Running the program: The program provides detailed on-screen instructions and help. ETCETERA can be run in any version of Windows except Windows 3**.**

Recalling results: Click on *"View"* in the top menu to display the current session's previous results

Pasting results: Results shown on the screen are automatically copied to the Windows clipboard, from which they can be pasted into a Microsoft Word or other text file at the site of the cursor (usually by pressing *Shift-Insert* or *Ctrl-V*. To ensure proper alignment of tabulated results, a Courier font should be used in the text file. If the current session's previous results are recalled (by clicking on "*View*"), text can be marked (drag the mouse over it with button pressed) and copied to the clipboard (by pressing *Ctrl-Insert* or *Ctrl-C*) for pasting elsewhere.

Adding comments: Click on "*Note*" in the top menu if you wish to add explanatory comments to be placed in the clipboard, saved, or printed with the results.

Saving results: By default, all results are saved in PEPI.TXT in the WinPepi folder, with a warning if it exceeds 500K. Results also go to PEPI.TMP (for display in the "*View"* option); this file may be overwritten unless it is renamed on quitting ETCETERA. Click on **"***Saving***"** (in the top menu) to see the default procedure or to change it, or to find a button that opens PEPI.TXT (which can also be accessed by clicking on "Results" in the Winpepi portal). [Results saved in earlier installations may be found in C:\PEPI.TXT]
TXT files can be combined with JOINTEXT (supplied with the Winpepi programs).

Printing results*:* Click on *"Print"*. If this fails, a simple solution is to paste the currently-shown results (which have automatically been copied to the Windows clipboard) into a Microsoft Word or other text program, and print from there. To ensure proper alignment of tabulated results, a Courier or similar font should be used in the text file. Results can also be printed from one of the files in which they are automatically saved, e.g. PEPI.TXT.

FINDING WHAT YOU WANT

FINDER.PDF (provided with this program) is an alphabetical index that identifies the modules (in all WinPepi programs) that deal with a specific procedure or kind of study. It is called up by pressing F9 or clicking on "*Finder*" in any WinPepi program, or on the FINDER icon, and can be printed for easy reference.

A DO-IT-YOURSELF THREESOME

1. The WinPepi suite of computer programs for epidemiologists, with their manuals. Can be downloaded free from www.brixtonhealth.com
2. "Research Methods in Community Medicine: Surveys, Epidemiological Research, Programme Evaluation, Clinical Trials" (J.H. Abramson and Z.H. Abramson), sixth edition. John Wiley & Sons, 2008.
3. "Making Sense of Data: A Self-Instruction Manual on the Interpretation of Epidemiological Data" (J.H. Abramson and Z.H.Abramson), third edition. Oxford: Oxford University Press, 2001.

HOW TO OBTAIN PEPI PROGRAMS

All WINPEPI (PEPI-for-Windows) and other PEPI programs can be downloaded free. The latest versions of WINPEPI programs – currently COMPARE2, DESCRIBE, ETCETERA, LOGISTIC, PAIRSetc, POISSON, and WHATIS – and their PDF manuals, can be downloaded from www.brixtonhealth.com. The latest release of Version 4 of PEPI, which contains over 40 DOS-based programs (which can be used in Windows) can be downloaded from www.sagebrushpress.com/pepibook.html
A printed manual is available for the DOS-based programs and WHATIS (Abramson and Gahlinger 2001).

Wilko C Emmens's XYgraph unit (version 2.2) creates the graphs displayed by this program.

A1. SIMPLE RANDOMIZATION (UNSTRATIFIED)

This module assigns subjects randomly to between 2 and 8 groups, each subject having an equal probability of assignment to each group. The groups are usually treatment or control groups in trials.

This simple randomization procedure may produce groups that (by chance) differ somewhat in size, especially if the number of subjects is small. Chance differences in the baseline characteristics of the groups are to be expected. To demonstrate that the groups exhibit random variation, and are not necessarily equivalent, the number and proportion of odd-numbered subjects in each group are reported (if there are 2 or 3 groups).

The candidates for selection must first be numbered in sequence, starting with 1 or any other number.

METHOD

The program uses a pseudo-random number generator described by Wichman and Hill (1985). Extensive statistical tests have demonstrated the statistical soundness of this algorithm, which derives each number in turn from three seed numbers (in the range 1 – 30,000) which it modifies for subsequent use. Initial values for the seed numbers are generated by Delphi's inbuilt random-number procedures, RANDOMIZE, which derives a preliminary seed from the system clock, and RANDOM,, which is used to generate three random numbers from which the required seed numbers are computed. Delphi's RANDOM procedure is augmented by an additional randomizing shuffle, using the algorithm of Bays and Durham, as described by Press *et al.* (1989: 215-217).

For simple randomization of subjects to G groups, the range R of random numbers ($0 < R < 1$) is divided into G equal fields, one for each group. A random number is selected for each subject in turn, and the assignment is determined by the field in which the random number falls. The probability of assignment to each group is $1 / G$.

A2. SIMPLE RANDOMIZATION OF SEPARATE STRATA

This module assigns subjects in different strata to between 2 and 8 groups, each subject having an equal probability of assignment to each group. The groups to which the subjects are allocated are usually treatment or control groups in trials.

The strata will usually reflect variables that it is believed may influence the outcome of the trial; the procedure prevents imbalance between the groups with respect to these variables. In a multicentre trial, each centre may be regarded as a stratum.

The maximum number of strata is 6; if there are more strata, module A1 should be applied separately in each stratum.

The candidates in each stratum must first be numbered in sequence, starting with 1 or any other number.

*This simple randomization procedure may produce groups that (by chance) differ somewhat in size, especially if the number of subjects in a stratum is small.

METHOD

The same method is used as in Module A1 (see above), applying it separately to each stratum.

A3 BALANCED RANDOMIZATION (UNSTRATIFIED)

This module assigns subjects randomly to between 2 and 8 groups, using a "biased coin" procedure that applies a constraint on the selection process to ensure that the groups to which the subjects are allocated are equal in size, or have any other required relative sizes (insofar as the total number of subjects permits this). The groups are usually treatment or control groups in trials.

Chance differences in the baseline characteristics of the groups are to be expected.

The candidates for selection must first be numbered in sequence, starting with 1.

METHOD

The program uses a pseudo-random number generator described by Wichman and Hill (1985). Extensive statistical tests have demonstrated the statistical soundness of this algorithm, which derives each number in turn from three seed numbers (in the range 1 –30,000), which it modifies for subsequent use. Initial values for the seed numbers are generated by Delphi's inbuilt random-number procedures, RANDOMIZE, which derives a preliminary seed from the system clock, and RANDOM,, which is used to generate three random numbers from which the required seed numbers are computed. Delphi's RANDOM procedure is augmented by an additional randomizing shuffle, using the algorithm of Bays and Durham, as described by Press *et al*. (1989: 215-217).

For balanced randomization of subjects to G groups, the range R of random numbers ($0 < R < 1$) is divided into G consecutive fields, one for each group. A random number is selected for each subject in turn, and the assignment is determined by the field in which the random number falls. The relative sizes of the fields are determined by the groups' quotas, i.e. the number of subjects that the groups require in order to meet their prespecified relative sizes. The quotas, and hence the probabilities of assignment, are calculated anew before the assignment of each subject, the probability that the next subject will be assigned to any specific group i being specified as A_i / N, where A_i is the number of additional subjects required to complete the quota for group i, and N is the number of subjects remaining to be assigned.

A4. BALANCED RANDOMIZATION OF SEPARATE STRATA

This module randomly assigns subjects in separate strata to between 2 and 8 groups, applying a constraint on the selection process to ensure that in each stratum the groups to which the subjects are allocated are equal in size, or have any other required relative sizes (insofar as the number of subjects in the stratum permits this). The groups to which the subjects are allocated are usually treatment or control groups in trials.

The strata will usually reflect variables that it is believed may influence the outcome of the trial. The procedure prevents imbalance between the groups with respect to these variables. In a multicentre trial, each centre may be regarded as a stratum.

The candidates for selection must first be numbered, the sequence starting with 1 in each stratum.

The maximum number of strata is 8; if there are more strata, module A3 should be applied separately in each stratum.

METHOD

The same method is used as in Module A3 (see above), applying it separately to each stratum.

A5. BALANCED RANDOMIZATION OF SUCCESSIVE BLOCKS

This module randomly allocates the subjects in successive blocks to between two and eight groups, applying a constraint on the selection process to ensure that in each block the same number of subjects (one or more) are allocated to each group. The groups to which the subjects are allocated are usually treatment or control groups in trials.

This method of randomization is appropriate in clinical trials in which the subjects are not known at the outset, but become available with the passage of time; in such studies, the procedure may serve to control for effects connected with the passage of time.

The blocks may be the same size as the number of groups, or a multiple of that number. The larger the blocks, the more difficult it becomes for clinicians to guess the assignment of the next candidate and to influence the assignment by deciding when to enter a subject into the study.

An option is offered for random selection of successive block sizes, in order to increase the probability that investigators will remain blind, and thus reduce possible bias (Efird 2011).

The subjects in each block must be numbered 1, 2, 3, etc.

The results may also be used if unequal assignments to groups are required. For example, if it is wished to have twice as many controls as treated cases, the module could be used to assign cases to groups A, B, and C, defining group A as the treatment group, and B and C (together) as the control group.

This blocked randomization procedure may be used in different strata, in order to prevent imbalance between the groups with respect to important variables. For this purpose, the module should be used repeatedly, for each stratum in urn. It may also be applied separately to each centre in a multi-centre trial.

METHOD

The same method of balanced randomization is used as in Module A3 (see above), but applying it to each block in turn.

A6. MINIMIZATION

Minimization is a method of balanced randomization, recommended for use in small trials (Altman 1991: 443-445, Altman and Bland 2006, Scott *et al*. 2002), whereby the assignment of each subject to a group is influenced by the distribution of selected prognostic categorical variables in the previously-assigned members of the groups.

Weighted randomization is used, in such a way as to bias the scales in favour of a decision that will minimize the differences between the groups with regard to these prognostic factors. This may make the findings of the study more persuasive, even in small studies, although it may reduce the power of conventional simple significance tests that do not include the prognostic factors in the analysis (Simon 1979, Scott *et al.* 2002).

Except in very large studies, minimization permits the control of more prognostic factors than stratification (Scott *et al.* 2002).

A separate decision must be made for each subject in turn, based on the prior findings in each group with respect to the prevalence of the selected prognostic factors. This – and especially the need to maintain a record of the prior findings in each group–- makes this a relatively inconvenient method, despite its effectiveness. The record provides a basis for the entry, for each subject except the first (who is allocated by a simple random decision.) of a "similarity score" for each group, based on the numbers of group members with the same attributes as the candidate. Each of the chosen attributes is treated separately for this purpose. For example, if the prognostic factors are sex, age, and the presence of diabetes, and the candidate is a diabetic man aged 35-44, and Group A already contains 8 men, 9 people aged 35-44, and 3 diabetics, the score for Group A at this stage is the sum of these numbers, i.e. 20 (8 + 9 + 3). The same weight is given to each prognostic factor. A random decision is then made, weighting the probabilities so that the candidate is most likely to be put in the group with the lowest score .

METHOD

The method proposed by Taves (1974), as described by Altman (1991: 443-445) and Scott *et al*. (2002), is based on the "similarity scores" (see above) that are entered .

The probabilities of assignment are determined in accordance with the similarity scores. Weighted randomization is used, setting the probability of assignment to the group with the lowest score at four times that of any other group. If there is a tie for lowest score, all groups that do not have the highest score are given a probability of assignment that is four times that of the group with the highest score. If there are ties both for the lowest score and for the highest score also, or if all scores are the same, an equal probability is set for each group.

A7. RANDOM SEQUENCING OF PROCEDURES

This module randomly determines the sequence of a set of two to eight procedures to which each subject in a study will be exposed.

It may be useful in circumstances where there is reason to believe that the effects of the procedures may be affected by their sequence of application. The procedures might, for example, be different treatment regimes whose effects it is wished to compare by applying them to the same subjects. Or, in an evaluative comparison of study methods, the procedures might be the examinations or interviews that it is wished to compare.

METHOD

The same method of balanced randomization is used as in Module A3 (see above), but applying it (for each subject) to the set of procedures under study.

B1. SIMPLE RANDOM SAMPLE, WITHOUT REPLACEMENT

This module selects a simple random sample of a specified size, or using a specified sampling fraction. Subjects are drawn one by one by the use of random numbers. Subjects who are selected are not returned to the pool of candidates, in order to ensure that they cannot be selected more than once. This is the kind of sample required in most studies (Cochran 1977: 18).

The selected subjects are listed both in the order of selection and in numerical order. The former listing may be useful in studies in which there is a possibility that the recruitment of subjects may be terminated prematurely because of lack of funds or other contingencies, since if candidates are recruited in the specified order the sample will be a random one (although not necessarily of the required size) at whatever point truncation occurs. It may also be useful in studies using *inverse sampling*, i.e. where the sample size is not determined in advance, but it is planned to continue sampling until a prespecified number of suitable study subjects have been identified (e.g., subjects who are revealed by a screening procedure to have evidence of a specific disease).

The candidates for selection must be numbered in sequence, starting with 1 or any other specified number.

A *stratified random sample* can be selected by choosing a separate simple random sample from each stratum in turn.. In each stratum, the candidates for selection must be numbered in sequence, starting with 1 or any other specified number. Different sampling fractions can be used in the different strata.

METHOD

The program uses a pseudo-random number generator described by Wichman and Hill (1985). Extensive statistical tests have demonstrated the statistical soundness of this algorithm, which derives each number in turn from three seed numbers (in the range 1 – 30,000), which it modifies for subsequent use. Initial values for the seed numbers are generated by Delphi's inbuilt random-number procedures, namely RANDOMIZE, which derives a preliminary seed from the system clock, and RANDOM, which is used to generate three random numbers from which the required seed numbers are computed. Delphi's RANDOM procedure is augmented by an additional randomizing shuffle, using the algorithm of Bays and Durham, as described by Press *et al*. (1989: 215-217).

The formula for each selection is

$$\text{trunc}(RM) + 1$$

where R is a random number in the range $0 < R < 1$
M = the number of candidates.

The same integer may be selected more than once, but previously selected numbers are filtered out.

B2. SIMPLE RANDOM SAMPLE, WITH REPLACEMENT

This module selects a simple random sample (in which each candidate has the same chance of inclusion) of a specified size, or using a specified sampling fraction.

Subjects are drawn one by one by the use of random numbers Subjects who are selected are returned to the pool of candidates, and may be selected again. A sample of this kind is occasionally required, since the formulae for the variances and estimated variances of estimates made from the sample are often simpler when sampling is with replacement (Cochran 1977: 18).

The selected subjects are listed in the order of selection. Repetitions are indicated by asterisks.

A stratified random sample can be selected by choosing a separate sample from each stratum. In each stratum, the candidates for selection must be numbered in sequence, starting with 1 or any other specified number. Different sampling fractions can be used in the different strata.

METHOD

The same method is used as in Module B1, except that previously selected numbers are not filtered out.

B3. TWO OR MORE SIMPLE RANDOM SAMPLES, WITHOUT REPLACEMENT

This module selects two to six simple random samples of specified sizes from a single pool of candidates. This may be useful in studies that aim to examine the reproducibility of findings, e.g. concerning the validity of a screening test, by comparing the findings in different samples. No subject is selected more than once. The candidates for selection must be numbered in sequence, starting with any chosen number .

METHOD

The same method is used as in Module B1, filtering out previously selected subjects, until the number in the combined samples has been chosen. The total group of selected subjects is then divided into consecutive samples, in accordance with the required sizes of the samples.

B4. RANDOM CHOICE OF ONE SUBJECT FROM EACH (EQUALLY-SIZED) SET

This module randomly selects one of the subjects in each of a number of sets of equal size (from 2 to 6). It might be used, for example, in a study in which cases of a disease are gradually accrued over time, and a subject is randomly selected from each successive set of patients.

METHOD

The program uses the pseudo-random number generator described above (see Module B1). In each set, the formula for the selection is

$$\text{trunc}(RM) + 1$$

where R is a random number in the range $0 < R < 1$

M = the size of the set.

B5. RANDOM SEQUENCE

This module arranges up to 5000 consecutive numbers in a random sequence.

It may be useful for determining the order of entry of subjects into a study, if there is a possibility that addition of subjects to the study may be terminated prematurely because of lack of funds or other contingencies; if candidates are added in the specified order the sample will be a random one (although not necessarily of the desired size) at whatever point truncation occurs. It may also be useful if *inverse sampling* is proposed, i.e. if the sample size is not determined in advance, but it is planned to continue sampling until a prespecified number of suitable subjects have been identified (e.g., subjects who are revealed to have a specific disease).

METHOD

The program uses the pseudo-random number generator described above (see Module B1). The numbers in the new sequence are selected one by one, without replacement, using the formula

$$\text{trunc}(RM) + 1$$

where R is a random number in the range $0 < R < 1$

M = the number in the sequence that have not yet been selected.

B6. TABLE OF RANDOM NUMBERS

This module displays as many tables of random numbers as are required.

Each table contains 144 numerals.

METHOD

The program uses the pseudo-random number generator described above (see Module B1).

B7. RANDOM DECISION (YES OR NO)

This module provides a random yes-no decision.

Each time the module is run it provides a random "yes" or "no" decision, equivalent to tossing a coin.

It may be of invaluable help to epidemiological researchers who are faced by critical decisions in their lives.

METHOD

The program uses the pseudo-random number generator described above (see Module B1).

C. MULTIPLE SIGNIFICANCE TESTS: ADJUSTED P VALUES

This module assists in the appraisal of multiple (simultaneous) significance tests performed on the same data. It may be used in situations where the probability of spuriously significant results (Type I errors) is elevated; for example, when there are a number of groups and each group is compared with each other group; when several groups are compared with the same control group; when several related hypotheses are tested in a comparison of two groups; or when the selection of associations for analysis is based not on prior hypotheses but on an examination of the data, and selected eye-catching differences are tested. Type I errors are particularly likely if data are "dredged" for statistically significant comparisons, without *a priori* hypotheses.

The P values are adjusted in such a way that whatever a*lpha* (critical P value) is used for appraising significance in individual tests, the probability of at least one spuriously significant result (Type I error) in the total set is no more than this *alpha*. Use is made of Holm's procedure (Holm 1979, Aickin and Gensler 1996) and Hommel's (Hommel 1988) and Finner's procedures (Finner 1990, 1993), which are based on the family-wise error-rate, and also of two procedures that use the false discovery rate (FDR) method (Benjamini and Hochberg 1995, Benjamini and Liu 1999).

Either the lowest P values in the set, or all the P values, must be entered, in any sequence. If all the values are not entered, the total number of tests in the set is required. This may be the number performed, or (if the tests were selected after examination of the data), the total number possible, including those not actually performed (Samuels 1991). In pairwise comparisons of N groups, for example, the number of possible tests is $N(N\text{-}1) / 2$.

Different P values occasionally yield identical adjusted P values. This is not an error.

Multiple comparison or simultaneous inference procedures adjust P values by taking account of the performance of multiple tests, to reduce the danger that associations will be reported as significant when they are flukes. Opinions on their use varies. 'It is to be hoped that they will become as much a part of accepted statistical practice as unadjusted P values are now,' says Wright (1992). Others consider them unnecessary, misleading, or inefficient (Rothman and Greenland 1998, Cole 1979, Perneger 1998) on theoretic grounds, because their use implies that the results of a test are interpreted differently according to how many other tests are performed, and because Type II errors may occur.

Bender and Lange (2001) say that 'different persons may have different but nevertheless reasonable opinions', but they 'prefer that data of exploratory studies be analyzed without multiplicity adjustment. "Significant" results … should clearly be labeled as exploratory results. To confirm these results the corresponding hypotheses have to be tested in further confirmatory studies.' Perneger (1998) concludes that multiple-comparison procedures make sense in only a

few situations. These include 'when searching for significant associations without pre-established hypotheses', as well as in repetitions of the same test in different strata or subsamples, and sequential testing of the results of a trial.

The program uses Holm's procedure (Holm 1979, Aickin and Gensler 1996) and Hommel's (Hommel 1988) and Finner's procedures (Finner 1990, 1993), which are based on the family-wise error-rate, and it also uses two procedures that employ the false discovery rate (FDR) (Benjamini and Hochberg 1995, Benjamini and Liu 1999). All these procedures are more powerful than the well-known Bonferroni procedure, which in effect adjusts the P value by multiplying it by the number of tests.

.

While different methods of handling multiple comparisons may be appropriate in different situations, Curran-Everett (2000) suggests that the false discovery rate (FDR) procedure described by Benjamini and Hochberg (1995), which is a "versatile, simple, and powerful approach", may be the best practical solution. The FDR is the expected proportion of false discoveries (false "statistically significant" results) among the discoveries. The two FDR methods used by the program are the "step-up" procedure described by Benjamini and Hochberg (1995), and a "step-down" procedure described by Benjamini and Liu 1999). The pros and cons of these alternatives are discussed by Benjamini and Liu (1999) and Benjamini *et al.* (2001).

For convenience, the observed *P* values are adjusted by multiplying them by *f* (the value of *f* depending on which multiple-comparisons procedure is used), instead of reporting that a specific observed *v*alue can be regarded as significant at the *alpha* significance level because it does not exceed *alpha / f*. An adjusted *P* value can then be regarded as significant (despite the multiple comparisons) if it does not exceed 0.05, 0.01, or any other chosen significance level. The adjusted P values for the FDR procedures are reported as <0.00001, <0.0001, <0.001, <0.01, <0.05, or >0.05.

In clinical trials in which multiple outcome measures are used, suggested solutions (instead of adjusting the P values) are selection of a single primary outcome measure, or creation of a global assessment measure (Feise 2002).

METHODS

Holm's procedure is a sequential one. Each P value in turn, starting with the lowest, is multiplied by *N - i*, where *N* is the total number of tests and *i* is the number of P values already adjusted. If an adjusted value is lower than a previous adjusted value, it is changed to the previous value, and if it exceeds 1 it is changed to 1.

Hommel's procedure is a more complicated sequential method, validated for independent tests; the program uses an algorithm provided by Wright (1992). If only some (i.e. the lowest) of the values are entered, the program makes the assumption that the missing values are evenly spaced between the highest value entered and 1. This is generally conservative, and may produce unduly high adjusted values for the higher values entered. As a precaution, Hommel's adjusted P is therefore not displayed for the three highest values entered, unless all values are entered.

For *Finner's procedure*, the P values are arranged in a sequence from lowest to highest (tied values are ranked consecutively), and the adjusted value of P_i (P value number *i* in the sequence) is computed as

$$1 - (1 - P_i)^{N/i}$$

where *N* is the total number of P-values in the set.
If an adjusted value is lower than a previous adjusted value, it is changed to the previous value, and if it exceeds 1 it is changed to 1.
The program uses an adaptation of a Fortran algorithm from MULTI (Version 2.0), by B.W. Brown and K. Russell (The University of Texas M. D. Anderson Cancer Center).

The *step-up false discovery rate (FDR*) procedure described by Benjamini and Hochberg (1995), is based on arrangement of the P values in a sequence from highest (i = N , i.e. the total number of P values in the set) to lowest (i = 1) (tied values receive consecutive ranks), and comparison with the appropriate FDR thresholds, calculated as the significance level (i.e. 0.05, 0.01, 0.001, and 0.0001, in turn) divided by f = N / i. This process identifies the points at which the next P value in the sequence is less than the FDR threshold for a more significant level, so that each P value can be reported as having an adjusted value in accordance with its relationship to the FDR thresholds (i.e., P < 0.00001, P < 0.0001, P < 0.001, P < 0.01, P < 0.05, or P > 0.05, as the case may be). For simplicity, adjusted P values are reported, computed as

P * *N / i*,

where *i* is the P value's rank in an ascending sequence of P values; but if an adjusted P-value is higher than the adjusted P value following it in this sequence, it is made equal to the following one.

For the *step-down false discovery rate (FDR)* procedure (Benjamini and Liu 1999), the program arranges the observed P values in a sequence from lowest (*i* = 1) to highest (*i = N*) and compares each observed value in turn (starting with the lowest) with the appropriate FDR thresholds, calculated by formula 2.1 of Benjamin and Liu.. This process identifies the points at which the next P value in the sequence is less than only the FDR threshold for a less significant level, so that each P value can be reported as having an adjusted value in accordance with its relationship to the FDR thresholds (i.e., P<0.00001, P<0.0001,bP < 0.001, P < 0.01, P < 0.05, or P > 0.05, as the case may be).

D. ASSESSMENT OF A SCALE

This module appraises the *internal consistency* and *discriminatory power* of a scale whose score is derived by summing the scores allotted to its constituent items. The items may relate to attitudes, practices, knowledge, the presence of symptoms, etc. They may all have Likert-like scores (e.g. 1, strongly agree; 2, agree; 3, undecided; 4, disagree; and 5, strongly disagree), or they may all be 'yes-no' (binary) items scored 1 or 0; in the latter case the total score is the number of 'yes' responses.

The program computes *Cronbach's alpha reliability coefficient*, the s*tandard error of measurement* and the 95% confidence interval for individual scores, and the *correlations* between each item and the total score and between each pair of items.

If the scale is based on 'yes-no' items, the program also computes approximate *tetrachoric correlations* between items, and appraises *conformity with a Guttman scale*. It reports the proportion of 'yes' responses for each item, the sequence in which the items would be placed in a Guttman scale, the percentage of individuals whose responses conform with perfect Guttman scale types, and error rates for each item. It computes a *coefficient of reproducibility*, and compares this with a coefficient of reproducibility by chance (CRC) and with the minimal marginal reproducibility (MMR). It also computes *coefficients of scalability*, performs a goodness-of-fit test, and provides a sensitivity analysis to assist in deciding whether the scale would be improved by the removal of specific items.

The program also computes *Ferguson's delta coefficient*, as a measure of the scale's discriminatory power, and performs a sensitivity analysis to appraise the effect of removing specific items. *Delta* is also computed for each scale item.

For a summated Likert scale, full data must be entered for each subject - i.e., each subject's score for each item in turn.

For a scale composed of 'yes-no' (1 or 0) items, three data-entry options are available - (a) separate entry of each subject's scores (which may be tedious if the sample is large), (b) entry of each pattern of responses and its frequency, or (c) entry of the frequency of each total score, and the frequency of 'yes' responses to each item. The frequencies required in options (b) and (c) must be determined in advance. If option (a) or (b) is used, the data can be pasted from a data file. If option (c) is selected, only the alpha coefficient is computed.

Items with only 'yes' responses and items with only 'no' responses must not be included in the scale, and missing values or missing-value codes are not permitted.

Cronbach's *alpha* coefficient

The *alpha* coefficient is a measure of the internal consistency, or 'internal-consistency reliability', of the scale, i.e., the extent to which the item responses correlate highly with each other. If the items were divided into two groups in every possible way, alpha would be the average correlation between the scores for the two 'split-halves' of the scale; it is essentially the square of the correlation between the observed score and the average score that would be obtained if the scale were applied an indefinite number of times (Cronbach 2004). A high value points to internal consistency, but does not necessarily mean that all the items measure the same dimension. If all items measure the same dimension, *alpha is* also the correlation between two applications of the scale (Heo et al. 2015)

A value of 0.7 is generally regarded as the lowest acceptable value, and a value of at least 0.8 is recommended. For clinical applications, a minimum of 0.9 has been recommended (Bland and Altman 1997).

For 'yes-no' items, an adjusted value of *alpha* is also computed, using Horst's formula (Guilford and Fruchter 1986: 429-430), which allows for differences between items in their 'difficulty' (i.e., in their proportions of 'Yes' responses). The usual formula for *alpha* assumes that the proportions of 'Yes' responses to the different items are similar.

If the items are all 'yes-no' items, *alpha* is equivalent to the Kuder-Richardson formula 20 (K-R 20) coefficient.

Standard error of measurement

The standard error of measurement, which is inversely related to alpha, is an estimate of error for use in interpreting an individual's score. It can be thought of as the standard deviation of the scores a subject would receive in repeated applications of the scale. The program uses it to estimate a 95% confidence interval for individual scores, on the assumptions (which are not necessarily true) that the error is the same at all levels of the score, and that the errors for any subject are normally distributed.

Correlations

Two sets of correlation coefficients are computed:
(a) correlations between each item and the total score (excluding the item from the total score). For yes-no items, coefficients of point biserial correlation between each item and the total score are calculated, with Henrysson's adjustment to compensate for the inclusion of the item in the total. The significance of the correlation is tested.
(b) correlations between each pair of items, and the mean inter-item correlation coefficient. If the scale is based on 'yes-no' items, approximate *tetrachoric correlations* between items are also computed; these provide an estimate of what the correlations would be if the distributions were not dichotomised, assuming an underlying distribution that is continuous and approximately normal; they are not computed if there is undue unevenness of the marginal totals (see Methods).

These coefficients permit the identification of items that it may be advisable to remove from the scale.

Guttman scale

A Guttman scale (or scalogram) is one whose items constitute a unidimensional series, such that a 'yes' response to any item predicts that the previous items in the series must also have 'yes' responses.

If the scale conforms with a Guttman scale, this suggests that the scale is a cumulative one (with a 'hierarchy' of responses), and that the items measure a single dimension. In most cases an individual's score (the number of 'Yes' responses) would be both a quantitative measure of this dimension and an indication of the specific pattern of responses.

To make the appraisal of conformity, the program first re-arranges the items in accordance with the frequency of positive responses, and defines Guttman scale types on the basis of this sequence. For example, for a three-item scale there are four acceptable patterns. When the items are arranged in order. from the one with the most 'Yes' responses to the one with the least, the perfect Guttman scale types are: 'Yes-Yes-Yes', 'Yes-Yes-No', 'Yes-No-No' and 'No-No-No'. Other patterns, e.g. 'Yes-No-No' , are non-scale types. The proportion of 'Yes' responses to each item and the sequence of the items in the scale are reported by the program.

Each individual's pattern of responses is then compared with the perfect Guttman scale type that has the same number of positive responses, and each discrepant response to a specific item is defined as an error. The percentage of individuals with perfect scale types is reported. The program also reports the error rate for each item, and its proportions of errors in what should be 'Yes' and 'No' responses. It has been suggested that the validity of a Guttman scale should be questioned if the errors for any item exceed 15%, or if over half the positive responses or over half the negative responses to any item are erroneous (Ford 1954: 294-295).

A coefficient of reliability is computed. This is the proportion of responses (in the total sample) that are not 'erroneous'; 0.9 is usually regarded as the minimal requirement for a satisfactory scale.

Since a high coefficient of reproducibility may be an expression of the overall distribution of responses to the various items, it is compared with a coefficient of reproducibility by chance (CRC), which is computed by first estimating the probability of each perfect scale type by multiplying the appropriate marginal probabilities, and then summing the probabilities of all perfect scale types (Riley 1963: 477). The program reports the absolute improvement achieved by the scale (the coefficient of reproducibility minus the CRC), and calculates a coefficient of scalability by dividing this by (1 - CRC). In a good Guttman scale, the coefficient of scalability should be well above 0.6. The program also computes alternative values of the absolute improvement and the coefficient of scalability, based on the minimal marginal reproducibility (MMR) instead of the CRC. The MMR is calculated by adding the marginal probabilities of all items (using the probability of either a positive or negative response, whichever is larger), and dividing the sum by the number of items (Nie et al. 1975: 528-533).

A sensitivity analysis is performed by recomputing the coefficients of reproducibility and scalability (based on the CRC) after removing each item in turn, in order to detect items whose removal would appreciably increase the scale's conformity with a Guttman scale.

The significance of the Guttman scale (Schuessler 1961) is appraised by an exact binomial goodness-of fit test that compares the observed coefficient of reproducibility with the computed coefficient of reproducibility by chance. A one-tailed mid-P value is displayed, expressing the probability of by chance obtaining a coefficient that is as high as, or higher than, the observed value. A low P value - say < 0.001 (Hofmann 1979) supports the possibility that the scale is a Guttman scale. It does not 'prove' that the scale is a Guttman scale; but Schuessler suggests that only if this P value is low should the various criteria listed above be applied.

Ferguson's delta coefficient

The scale's *discriminatory power* can be measured with Ferguson's *delta* coefficient, which ranges from 0 if all subjects have the same scale score to 1 if subjects are equally divided among all possible scale scores. A scale may be considered discriminating if *delta* is above 0.9.

Delta is computed for scales composed of 'yes-no' items scored 1 or 0, or of items that have Likert-like scores that all have the same range (e.g., 0, 1, 2, or 1, 2, 3, 4, 5).

Delta is computed for the total scale, for the total scale excluding each item in turn, and for each separate item.

Note that the removal of uncorrelated but valid items may reduce the scale's discriminatory power, whereas heterogeneity of the items may increase discriminatory power at the expense of internal-consistency reliability.

METHODS

The program can deal with scales containing up to 60 items (if they all have single-digit scores), and up to 4000 subjects (if individual subjects are entered) or up to 4000 patterns of 'yes-no' responses. Because of the limited size of the data entry box, the maximum number of items is 37 if all items have two-digit scores, and 46 if half have two-digit scores.

Cronbach's alpha coefficient and related statistics:

The formula for *alpha* ((Guilford and Fruchter 1986: 428) is

$$[k / (k - 1)] (1 - \sum(s_i^2 / s_t^2)$$

where k = number of items in the scale
s_i = standard deviation of scores for item i
s_t = standard deviation of total scores

For a scale composed of 'yes-no' items, $\sum p_i q_i$ is substituted for s_i^2 in the above formula
Where p_i = proportion of 'yes' responses to item i
$q_i = 1 - p_i$
this is the Kuder-Richardson formula 20 (Guilford and Fruchter 1986: 427-428.

Horst's modified Kuder-Richardson formula, adjusting for differences in item difficulty (Guilford and Fruchter 1986: 429-430), is:

$$[(s_i^2 - \sum p_i q_i / \{(s_m^2 - \sum p_i q_i)] (s_m^2 / s_t^2)$$

where $s_m^2 = 2^* \sum R_i p_i - T(1 + T)$
T = mean score
R_i = rank position of item *i*, (the item with a lowest pi being ranked 1)

The standard error of measurement (SE_M) is $s_t\sqrt{[(1 - alpha)]}$

The 95% confidence interval for individual scores is (score - 1.96SE_M) to (score + 1.96SEM).

Correlations

If the scale is composed of 'yes-no' items, coefficients of point biserial correlation between each item and the total score are calculated, with Henrysson's adjustment (Guilford and Fruchter 1986: 466) to compensate for the inclusion of the specific item in the total score.

In some unusual circumstances the program skips the calculation of coefficients, especially adjusted coefficients, because of computational difficulties.

Approximat*e tetrachoric correlation coefficients* are calculated by the formula proposed by Edwards and Edwards 1984):

$$r = (OR^{pi/4} - 1) / (OR^{pi/4} - 1)$$

where OR = odds ratio
a and *d* = numbers of concordant pairs
b and *c* = numbers of discordant pairs

This simple method, which was used by Stata until recently, provides an approximation that is acceptable in many situations (Digby 1983, referring to an almost identical formula, with ¾ instead of pi/4) but can be very inaccurate (Uebersax 2000). V. Wiggins, of the Stata Corporation, in a reply cited by Gunther and Hofler (2006), says that the approximation works well when the marginals in both directions are above 10%. ETCETERA does not display the coefficient unless this condition is met.

Conformity with a Guttman scale

The methods are explained above.

Ferguson's *delta* coefficient

Delta is computed by a formula (*deltaG*) that is applicable both to 'yes-no' items and to items with more responses (Hankins 2007):

$$DeltaG = [(1 + k(m - 1)](n^2 - S) / n^2k(m - 1)$$

where n = sample size
k = number of items in scale
m = number of possible responses (from zero to top score) to each item
f_i = frequency of scale score *i*
$S = \sum f_i^2$

E. APPRAISAL OF STATISTICAL SYNERGISM

This module appraises *statistical synergism* between two binary ("yes-no") variables, *A* and *B*, with respect to a binary "outcome" variable *C*. Statistical synergism (or antagonism) does not necessarily mean biological (causal) interaction. The computation is based on comparisons of risk ratios or odds ratios.

The program provides *tests for synergy* on additive and multiplicative scales, and several *measures of synergy* on both scales - the interaction contrast (*IC*), the interaction contrast ratio (*ICR*, also called *RERI*, the relative excess risk due to interaction ,(the attributable proportion due to interaction (*AP*), the attributable proportion due to interaction among cases attributable to the combined factors (*AP**, or *APstar*), Rothman's synergy index (*SI*), and the synergy factor (*SF*). Confidence intervals for the *ICR*, *AP*, *SI* and *SF* measures, and the statistical tests, are provided if the frequencies in the 4x2 table (see below) are entered.

Three modes of data entry are offered. First, the risks of *C* can be entered, i.e. the risk of *C* when only *A* is present, and when only *B* is present, and when both *A* and *B* are present, and when neither *A* nor *B* is present. Risk ratios are then computed for use in the analysis.

Secondly, a 4x2 table can be entered, showing the numbers with and without *C* (the "outcome") when *A* is present and *B* absent, when *B* is present and *A* absent, when both are present, and when neither is present. The main analysis is then based on risk ratios, but measures based on odds ratios, which may differ from those based on risk ratios, are also provided. In case-control studies with unequal sampling fractions for cases and controls, the risk ratios are derived from ancillary information on the ratio of these fractions, or roughly estimated from the prevalence of cases in the population.

Thirdly, odds ratios can be entered, and the program will then calculate measures based on odds ratios onlys.

Tests for synergy

Statistical tests for synergy, or (more accurately) tests for departure from an interaction-free additive model and from an interaction-free multiplicative model (de Gonzalez and Cox 2005) are performed if the frequencies in the four-by-two table are entered. The tests permit an assessment of whether the data are consistent with neither, one, or both of the two models, namely additive with no interaction, and multiplicative with no interaction. It is assumed that the disease (or other outcome variable) is rare.

Synergism on an additive scale

Five measures of s*ynergism on an additive scale* (Rothman 1986) are computed, based on comparisons of the joint effect of *A* and *B* with the sum of their separate effects. They are

(a) *IC* (the interaction contrast), which is the excess risk due to interaction;

(b) *ICR* (the interaction contrast ratio), which is also called the *RERI*, the relative excess risk due to interaction (the excess risk due to interaction, relative to the risk in the absence of A and B);

(c) *AP* (the attributable proportion), which, if positive, is the proportion of cases attributable to the interaction of A and B;

(d) *AP** (*APstar)*, which (if positive) is the proportion of cases attributable to the interaction of A and B, among subjects exposed to both A and B; and

(e) Rothman's *SI* (synergy index), which is the excess risk from exposure to both A and B when there is interaction, relative to the excess risk from exposure to both A and B in the absence of interaction.

The first four of these have a zero value if there is no additive interaction, whereas the null value of SI is 1. If the measures exceed their null values, the possibility of biological (causal) synergism may be considered. Statistical synergism on an additive scale – i.e., evidence that the joint effect is greater than the sum of the separate effects (rather than their product) – is generally regarded as the minimum requirement before considering biological synergism. If the measures are below their null values, this indicates reduced additivity, but is not evidence of antagonism; the possibility of biological antagonism may be considered if the joint effect is smaller than the separate effects of both *A* and *B*.

If the four-by-two table showing numbers with and without *C* is entered, 90%. 95%, and 99% confidence intervals for *ICR*, *AP*, and *SI* are estimated, using the "MOVER" procedure (method of variance estimates recovery) described by Zou (2008), whose simulation studies have demonstrated its appropriateness. Also, a significance test is provided for departure of *SI* from its null value of 1. This test, which does not always conform with the confidence intervals, assumes that the sample sizes are reasonably large.

Synergism on a multiplicative scale

A *measure of multiplicative interaction* based on risk ratios is also computed, based on a comparison of the joint effect of *A* and *B* with the product of their separate effects, as well as the *synergy factor* (*SF*) suggested by Cortina-Borja *et al.* (2009), which is based on odds ratios. The null values are 1.

Risk ratios

The risk ratios required for calculating the measures of interaction can be computed from the risks of C – i.e., its risk when only A is present, when only B is present, when both A and B are present, and when neither A nor B is present.

If these risks are not entered, the risk ratios can be computed from the four odds (in favour of C) – i.e., the odds when A is present and B absent, when B is present and A absent, when both are present, and when both are absent. This computation requires the frequencies of C and its absence, in each of these circumstances. Since the odds in a case-control study are affected by the sampling probabilities for cases (subjects with C) and controls (subjects without C), the calculation takes account of the ratio (if it is entered) of these sampling fractions. If the ratio of sampling fractions is not entered, the program can roughly estimate it from the overall rate or proportion of cases in the population studied..

Odds ratios versus risk ratios

The use of odds ratios rather than risk ratios tends to exaggerate the interaction (Zou 2008), as is obvious in the program outputs in which both are used. It may yield results that diverge appreciably (especially for the *ICR* and *SI* measures) from those based on risk ratios, their divergence varying with the baseline risk and the magnitude of interaction (Kalilani and Atashili 2006). For more than additive interaction, the difference is more pronounced for the *ICR* and *SI* measures, and for less than additive interaction it is more marked for *AP*. Even when the outcome (C) is rare, the use of odds ratios may point to interaction (additive or multiplicative) in instances where the use of risk ratios would indicate the absence of interaction (Campbell *et al*. 2005).

METHODS

Synergy tests

The tests for departure from the additive or multiplicative model (using risk ratios) are described by De Gonzalez and Cox (2005, formulae 6 and 16). They are performed only if the risk when A and B are present exceeds the expected risk according to the relevant model. In a case-control study, for the purpose of these tests the odds estimates used for this purpose are first adjusted by dividing them by the ratio of the sampling fractions used for cases and controls. If this ratio is not entered, it is estimated roughly by dividing the observed ratio of cases to non-cases by the ratio of cases to non-cases in the population .

A test based on odds ratios, for the analysis of case-control data, uses formula 22 of De Gonzalez and Cox (2005).

The significance of *SI* and *SF* is appraised by *z* tests (Hogan *et al.* 1978 and Cortina-Borja *et al.* 2009, respectively) if the outcomes (i.e., the numbers with and without C) are entered. One-tailed P values are displayed.

The various tests and confidence intervals may not be consistent with one another.

Tests for interaction

The tests for departure from the additive or multiplicative model are described by de Gonzalez and Cox (2005, formulae 6 and 16). They are performed only if the rate in the population is entered, and the risk when A and B are present exceeds the expected risk according to the relevant model. In a case-control study, for the purpose of these tests the reported numbers of cases are first divided by the ratio of the sampling fraction for cases to the sampling fraction for controls; if this ratio is not entered, it is estimated by dividing the observed case-control ratio by the rate in the population (Rothman and Greenland 1998: p. 418). One-tailed P values are displayed.

Measures of additive interaction:

The following formulae for *IC*, *ICR*, *AP*, and *SI*, based on risk ratios (see below), are provided by Kalilani and Atashili 2006 (formula 1-4) ; but note that the correct denominator in the formula for *SI* is

$(RR_{10} - 1) + (RR_{01} - 1)$, and not
$(RR_{10} - 1)(RR_{01} - 1)$, as printed.

$IC = R_{11} - R_{10} - R_{01} + R_{00}$
$ICR = IC / R_{00}$
$AP = IC / R_{11}$
$SI = (RR_{11} - 1) / [(RR_{10} - 1) + (RR_{01} - 1)]$

The formula for *AP** (APstar) (based on Rothman 1986: 322 and 325) is:
$AP^* = AP / [(RR_{11} - 1) / RR_{11})]$

In analyses based on odds ratios, the risk ratios in the above formulae are replaced by odds ratios.

Confidence intervals are estimated by the formulae provided by Zou (2008), and tally very closely with the results computed by a spreadsheet provided by Zou.

Risk ratios

These formulae are based on risk ratios, derived from the risks of C:
$RR_{10} = R_{10} / R_{00}$
$RR_{01} = R_{01} / R_{00}$, and
$RR_{11} = R_{11} / R_{00}$
where R_{10} = risk of C when only A is 'yes'
R_{01} = risk of C when only B is 'yes'
R_{11} = risk of C when both A and B are 'yes'
R_{00} = risk of C when both A and B are 'no' '

If the risks are not entered, the risk ratios are calculated from odds estimates (the odds in favour of C under the above four conditions) derived from the data, using Kalilani and Atashili's formulae 8-10 (which are based on the odds estimates and their ratios), and these are employed in the additive-interaction formulae; in effect, Kalilani and Atashili' s formulae 12, 14, and 15 for *ICR*, *AP*, and *SI* are used. In a case-control study, the odds estimates used for this purpose are first adjusted by dividing them by the ratio of the sampling fractions used for cases and controls. If the ratio of sampling fractions is not entered, it is estimated by dividing the observed case-control ratio by the rate or proportion of cases in the population (Rothman and Greenland 1998: p. 418), if this is entered.

Index of multiplicative interaction:

The index of multiplicative interaction is based on the risks or risk ratios. The formula is : \

RR11 / (RR10 * RR01) (Campbell *et al.* 2005),

which is equivalent to (R11 * R00) / (R10 * R01).

The formula for the *synergy factor*, based on odds ratios, is

OR11 / (OR10 * OR01) (Cortina-Borja *et al*. 2009)

F1. CORRELATION COEFFICIENT: TESTS, CONFIDENCE INTERVALS, UNBIASED ESTIMATES

This module provides tests and other procedures applicable to a Pearson's correlation coefficient – a simple correlation coefficient (e.g. r_{AB}), a partial correlation coefficient (e.g.$r_{AB.C}$ or $r_{AB.CD}$), or a multiple correlation coefficient (e.g. $R_{A.BCDE)).}$

If a *simple correlation coefficien*t is entered, the program computes its *significance* in comparison with zero and (optionally) in comparison with any other selected value, provides an unbiased estimate of the *population correlation coefficient*, and estimates its 90%, 95%, and 99% *confidence intervals.*

If a *partial or multiple correlation coefficient* is entered, the program computes its *significance* in comparison with zero, and provides an unbiased estimate of the *population correlation coefficient*.

If a *simple or multiple correlation coefficient* is entered, the program displays the *coefficients of determination, nondetermination, and alienation* and the *index of forecasting efficiency.*

Besides the coefficient, the size of the sample must be entered. If the coefficient is a partial correlation coefficient, the total number of variables is required; and if it is a multiple correlation coefficient, the number of independent variables must be entered.

The *population correlation coefficient* is an unbiased estimate of the correlation in the population represented by the sample studied. It is valid only if the variation between individuals in the sample and in the population are comparable (Oldham *et al.* 1992).

The *coefficient of determinatio*n (based on a simple correlation coefficient) is the proportion of variability in one of the variables that can be accounted for by its correlation with the second variable. The *coefficient of multiple determination* (based on a multiple correlation coefficient) is the amount of variability in the dependent variable that is explained by the correlation with the other variables. The *coefficient of nondetermination* is the amount of variability that is not explained by the correlation, and the *coefficient of alienation* indicates the degree of lack of relationship (Guilford and Fruchter 1986).

The *index of forecasting efficiency* is the estimated percentage reduction in errors of prediction by reason of knowledge of the correlation.

METHODS

Note
Some procedures are omitted if the number of variables is too large for the sample size.

Significance tests

Comparison with zero correlation:
The significance of a *simple correlation coefficient* is tested by the formula

$$t = [r\sqrt{(n - 2)} \ / \sqrt{(1 - r^2)}, \text{ with } (n - 2) \text{ degrees of freedom}$$

where r = correlation coefficient
n = sample size

The significance of a *partial correlation coefficient* is tested by an F test with 1 and (n - 3) degrees of freedom (Blalock 1979: 496):

$$F = r^2 \ (n - v - 1) \ / (1 - r^2)$$

where v = total number of variables.
The significance of a *multiple correlation coefficient*, *R*, is tested by an *F* test with v and $(n - v - 1)$ degrees of freedom (Blalock 1979: 494; Howell 1997: 522):

$$F = R^2 \ (n - v - 1) \ / v(1 - R^2)$$

where v = number of independent variables.

*Comparison with a nonzero correlatio*n:
If the sample size is 30 or more, significance is tested by the formula (Sokal and Rohlff 1981: 517):

$$t = (T1 - T2)\sqrt{(n - 3)}$$

where *T1* and *T2* = *z* transformations of the two values of *r*
If the sample size is less than 30, significance in comparison with a nonzero correlation is tested by the formula (Sokal and Rohlff 1981: 518):

$$t = (H1 - H2)\sqrt{(n - 1)}$$

where *H1* and *H2* = Hotelling's modified *z* transformations of the two values of *r*.

Note
Tests using Hotelling's transformation should be regarded as approximate if the sample size is less than 25 (Sokal and Rohlf 1981: 519).

Population correlation coefficient

For a *simple correlation coefficient*, the formula is

$$\sqrt{\{[r^2 * (n - 1) - 1] / (n - 2)\}}$$

For a *partial correlation coefficient*, the formula (Croxton & Cowden 1939: 775) is

$$\sqrt{\{r^2)(n - 1) / (n - v - 1)]\}}$$

For a *multiple correlation coefficient*, *R*, the formula (Howell 1997: 521) is:

$$\sqrt{\{1 - [(1 - R^2)(n - 1) / (n - v - 1)]\}}$$

where v = number of independent variables.

Note: If the correlation is very low, the number whose square root is taken as the population correlation coefficient may be negative, and the population correlation coefficient is then arbitrarily displayed as 0 (Croxton and Cowden 1939: 679).

Confidence intervals

For a *simple correlation coefficient*, confidence intervals for the population correlation coefficient (Altman and Gardner 2000: 89) are estimated as

$$[\exp(2 * F) - 1] / [\exp(2 * F) + 1] \text{ to } [\exp(2 * G) - 1] / [\exp(2 * G) + 1]$$

where $F = Z - A / \sqrt{(n - 3)}$

$G = Z - A / \sqrt{(n - 3)}$

$Z = \log((1 + r) / (1 - r)) * 0.5$

$A = 1.645$, 1.96, or 2.576 for 90%, 95%, or 99% confidence intervals respectively.

The same procedure is used for a *partial correlation coefficient*, with the following changes (Blalock 1979: 496):

$F = Z - A / \sqrt{(n - v - 1)}$

$G = Z - A / \sqrt{(n - v - 1)}$

v = total number of variables

The same procedure is used for a *multiple correlation coefficient*, with the following changes (Blalock 1979: 496):

$F = Z - A / \sqrt{(n - v - 2)}$

$G = Z - A / \sqrt{(n - v - 2)}$

v = number of independent variables

Coefficients of determination, nondetermination and alienation

The *coefficient of determination* [or *multiple determination*] is r^2, the *coefficient of nondetermination* is $1 - r^2$, and the *coefficient of alienation* is $\sqrt{(1 - r^2)}$,

where r = a simple correlation coefficient or the unbiased estimate of the multiple R in the population.

Index of forecasting efficiency

This index = $100(1 - \sqrt{(1 - r^2)}$.

F2. APPRAISAL OF INDEPENDENT CORRELATION COEFFICIENTS

This module appraises two or more Pearson's correlation coefficients that are based on different unmatched samples.

It computes *significance* and 95% *confidence intervals* for each coefficient, and estimates the *common correlation coefficient* – i.e. the value of the coefficient in the population represented by the samples, with its confidence intervals (an estimate that is not valid if there is significant heterogeneity).

If only two coefficients are entered, the program tests their difference, and if more than two are entered, it performs a *heterogeneity test*.

Approximate 95% *confidence intervals for the differences between coefficients* are computed.

If more than two coefficients are entered, *pairwise comparison*s are performed, using a Tukey-type multiple-test procedure. As an optional alternative, one of the coefficients can be designated as a control value, with which each of the others is compared, using a Dunnett-type multiple-test procedure.

METHODS

Significance of coefficients

The significance of each coefficient, in comparison with zero correlation, is tested by Zar's formula 19.4 (Zar 1998: 381) if the sample size is 30 or more:

$t = [r\sqrt{(n - 2)} \ / \ \sqrt{(1 - r^2)}$, with $(n - 2)$ degrees of freedom

where r = correlation coefficient

n = sample size

If the sample size is less than 30, a test based on Hotelling's modified *z* transformation is used (Zar 1984: 392, Sokal and Rohlf 1981: 587)

$z = H \sqrt{(n - 1)}$

where H = Hotelling's modified *z* transformation of r

$= T - (3T + r) / 4n$

T = *z* transformation of r

$= 0.5\ln[(1 + r) / (1 - r)]$

Note: Tests using Hotelling's transformation should be regarded as approximate if the sample size is less than 25 (Sokal and Rohlf 1981: 519).

Confidence intervals for coefficients

A 95% confidence interval for the population correlation coefficient (Altman and Gardner 2000: 89) is estimated as

$$[\exp(2 * F) - 1] / [\exp(2 * F) + 1] \text{ to } [\exp(2 * G) - 1] / [\exp(2 * G) + 1]$$

where $F = Z - A / \sqrt{(n - 3)}$

$G = Z - A / \sqrt{(n - 3)}$

$Z = \log((1 + r) / (1 - r)) * 0.5$

$A = 1.96$

Confidence intervals for differences between coefficients

Approximate confidence intervals for the differences between coefficients are computed by the modified asymptotic methods described by Zou (2007), using formula 15. They are based on the confidence intervals of the separate coefficients. There may be discrepancies between the confidence intervals and the results of the significance tests.

Common correlation coefficient

The common correlation coefficient (Zar 1998: formula 19.32, p. 390) is estimated by calculating its *z* transformation, z_c, as

$$\sum[(n_i - 3)z_i] / \sum(n_i - 3)$$

where n_i = size of sample *i*

z_i = z transformation of coefficient *i*

and then converting z_c to the corresponding correlation coefficient, r_c:

$$r_c = [\exp^{2\,zc} - 1] / [\exp^{2\,zc\,+} 1]$$

Paul's formulae (Paul 1988) are used as well; these are said to provide better estimates if the coefficient is less than about 0.5:

For two coefficients, this is Zar's formula 19.26 (Zar 1998: 388):

$$z_c = [(n_1 - 1)z'_1 + (n_2 - 1)z'_2] / [(n_1 - 1) + (n_2 - 1)]$$

and z_c is then converted to the corresponding correlation coefficient, r_c:

$$r_c = [\exp^{2\,zc} - 1] / [\exp^{2\,zc\,+} 1]$$

where, for each value of *z*,

$$z' = z - (3z + r) / [4 (n - 1)]$$

For three or more coefficients, $(n_i - 3)$ is replaced by $(n_i - 1)$ in formula 19.32 (see Zar 1998: 392).

The *significance* of the common correlation coefficient (in relation to zero) is computed by Paul's formula (Zar 1998: 390, formula 19.35):

$$\text{chisq} = \sum \{[n_i(r_i - r_c)^2] / (1 - r_i.r_c)^2$$

with $k - 1$ degrees of freedom,

where r_i = coefficient *i*

r_c = common correlation coefficient

k = number of coefficients.

An approximate 95% *confidence interval* is computed by the formula used for single correlation coefficients (see above), using the combined sample sizes as *n*.

Comparison of correlation coefficients

For two correlation coefficients, Zar's formula 19.21 (Zar 1998: 386) is used:

$$Z = (z_1 - z_2) / \sqrt{[1 / (n_1 - 3) + (1 / \{n_2 - 3\})]}$$

where z_1, z_2 are the z transformations of the coefficients

n_1, n_2 are the sizes of the two samples.

For three or more correlation coefficients, the *heterogeneity test* (Zar 1998: 390, formula 19.31) is:

$$chisq = \sum[(ni - 3)zi^2] - \sum[(ni - 3)zi]^2 / \sum(ni - 3)$$

Multiple pairwise comparisons of correlation coefficients

If from 3 to 40 coefficients are entered, multiple pairwise comparisons are performed by a Tukey-type test (Zar 1996: 393, formula 19.36) and appraised by referring to critical values of the *Q* distribution for P = 0.001, 0.01, and 0.05 (Zar 1996: Table B.5). Gaps in the table of critical values are filled by harmonic interpolation.

As an optional alternative (if from 3 to 20 coefficients are entered), multiple comparisons with a single specified 'control' coefficient are performed (Zar 1996: 394, formula 19.39) and appraised by reference to critical values for Dunnett's test for (two-tailed) P = 0.01 and 0.05 (Zar 1996: Table B.7; Dunnett 1964: Tables II and III,). Gaps in the tables of critical values are filled by harmonic interpolation.

F3. APPRAISAL OF CORRELATION COEFFICIENTS BASED ON THE SAME SAMPLE

This module appraises two or more Pearson's correlation coefficients that are based on the same sample.

The coefficients to be appraised (up to 15) must be entered in the left-hand data box, specifying the variables whose correlation they measure. For this purpose, the variables should be allocated consecutive numbers – 1, 2, 3 etc. Unless these coefficients comprise a complete matrix, the coefficients for all other combinations of the specified variables should be entered in the right-hand data box.

The program computes *significance* and 95% *confidence intervals* for each coefficient listed in the left-hand box, and it tests the *difference between each pair* of these coefficients. Approximate 95% *confidence intervals for the differences between coefficients* are computed. The program also performs *heterogeneity tests* for sets of three or more coefficients.

Pairwise comparisons

Two tests are used for comparisons of correlations that overlap (i.e., those that have a variable in common): the tests described by Meng *et al*. (1992) and by Dunn and Clark (1969). A simulation study led Hittner *et al*. (2003) to recommend Dunn and Clarks's test for its overall statistical properties. The method of Raghunathan *et al.* (1996) is used if the correlations do not overlap. Approximate 95% confidence intervals for the differences between coefficients are computed; the intervals are not always consistent with the results of the significance tests

Since a large number of pairwise tests may be performed, there is a possibility that apparently significant findings may be flukes. In addition to the P values estimated by the tests, the program therefore provides adjusted P values that take account of the performance of multiple tests. For this purpose, use is made of Finner's procedure (Finner 1990, 1993), which is more powerful than the well-known Bonferroni method.

Heterogeneity tests

If more than two coefficients are entered in the left-hand box, Raghunathan's approximate test (Raghunathan 2003) is applied. This appraises the heterogeneity of all these coefficients (irrespective of whether or not they have variables in common), while controlling for other correlations (if any) between the specified variables.

In addition, the methods of Meng *et al*. (1992) is used to compare any sets of three or more coefficients (among those entered in the left-hand box) that have a variable in common.

METHODS

Significance of coefficients

The significance of each coefficient, in comparison with zero correlation, is tested by Zar's formula 19.4 (Zar 1998: 381) if the sample size is 30 or more:

$t = [r\sqrt{(n - 2)} \ / \ \sqrt{(1 - r^2)}$, with $(n - 2)$ degrees of freedom

where r = correlation coefficient

n = sample size

If the sample size is less than 30, a test based on Hotelling's modified z transformation is used (Zar 1984: 392, Sokal and Rohlf 1981: 587)

$z = H \sqrt{(n - 1)}$

where H = Hotelling's modified z transformation of r

$= T - (3T + r) / 4n$

T = z transformation of r

$= 0.5\ln[(1 + r) / (1 - r)]$

Note: Tests using Hotelling's transformation should be regarded as approximate if the sample size is less than 25 (Sokal and Rohlf 1981: 519).

Confidence intervals for coefficients

A 95% confidence interval for the population correlation coefficient (Altman and Gardner 2000: 89) is estimated as

$[\exp(2 * F) - 1] \ / \ [\exp(2 * F) + 1]$ to $[\exp(2 * G) - 1] / [\exp(2 * G) + 1]$

where $F = Z - A / \sqrt{(n - 3)}$

$G = Z - A / \sqrt{(n - 3)}$

$Z = \log((1 + r) / (1 - r)) * 0.5$

$A = 1.96$

Confidence intervals for differences between coefficients

Approximate confidence intervals for the differences between coefficients are computed by the modified asymptotic methods described by Zou (2007), using applications of formulae 13 and 14 to (respectively) overlapping correlations, i.e. those that have a variable in common (Example 2), and nonoverlapping correlations (Example 3). The intervals are based on the confidence intervals of the separate coefficients, and take account of the dependencies between the correlations that are compared.

Pairwise comparisons

Formula 1 of Meng *et al*. (1992) and Dunn and Clark's test are used for pairwise comparisons of *overlapping coefficients*. Formulae for both tests are cited by Hittner e*t al.* (2003) . For *nonoverlapping coefficient*s, the program applies the ZTP (modified Pearson-Filon) procedure described by Raghunathan *et al*. (1996) (formula 3), with an approximate method of adjusting for nonindependence (formula 6). If over three pairwise comparisons are done, the Dunn-Clark and Raghumnathan tests are used, and the P values are supplemented by values adjusted for multiple testing, using the procedure described by Finner (1990, 1993).

Heterogeneity tests

The formula for the test statistic for Raghunathan's approximate test is expressed in formula 1 of Raghunathan (2003). P values are based on the chi-square distribution. If there are fractional degrees of freedom, P values are estimated approximately, using harmonic interpolation between the integers (Zar 1998L: App10).

F4. COMPUTATION OF PARTIAL AND MULTIPLE COEFFICIENTS

This module computes *partial and multiple correlation coefficients* based on the correlations (in the same sample) between up to nine variables. Either Pearson's correlation coefficients or rank correlation coefficients (Spearman's or Kendall's) may be entered (Lehmann 1977).

The module provides both *first-order partial correlation coefficients* (e.g. *r*12.4), and, optionally, *second-order partial correlation coefficients* (e.g. *r*12.45), and the s*quared partial correlation coefficient*s.

For each first-order partial correlation coefficient, the extent to which the third variable affects the correlation is examined by estimating *95% confidence intervals for the difference between the simple and first-order coefficients.*

If correlations between three or four variables are entered, it can also compute *multiple correlation coefficients (*e.g. *R*1.24) and their squares (*coefficients of multiple determination, R^2*), with *unbiased estimates of the multiple correlation coefficient in the population*.

If the sample size is entered, the significance of the correlations is tested, and confidence intervals are estimated for partial Pearson's correlation coefficients. If the sample size varies (because of missing data), entry of the smallest size will provide conservative tests and intervals.

Partial correlation coefficients

First-order partials (e.g. *r*12.4) express the linear correlation between two variables when a third variable is controlled, and *second-order partials* (e.g. *r*12.35) express the linear correlation between two variables when two others are controlled.

Optionally, the significance of the coefficients (in comparison with zero) is tested, and 95% confidence intervals are estimated. Caution should be used in interpreting the significance tests for partial correlation coefficients (Siegel and Castellani 1988: 261) since if there are many such tests there is a considerable risk of obtaining spurious significance. Since the standard error of the z transform of Spearman's *rho* is 1.03 times that of the standard error of Pearson's *r,* and the standard error of Kendall's *tau* is 0.66 times that of Pearson's *r* (Fieller *et al*. 1957, 1961), separate tests are conducted for partial rank correlation coefficients.

The *squared partial correlation coefficients* are also displayed. These reflect the percent of unexplained variance in the dependent variable that is explained by adding the control variable or variables. The square of r12.4 can be interpreted as the percent of the variance in variable 1 not accounted for by variable 2, that is accounted for by variable 4.

For each first-order partial correlation coefficient, the extent to which the third (control) variable affects the correlation is examined by estimating *95% confidence intervals for the difference between the simple (zero-order) and first-order coefficients.* If the confidence interval does not straddle zero, this points to a significant effect ($P < 0.05$). This difference expresses the extent to which the correlation can be attributed to the control variable or (if the partial coefficient is larger than the simple coefficient) the influence of the control variable as a suppressor variable.

Multiple correlation coefficients

Multiple correlation coefficients (e.g. *R*1.24) measure the combined influence of two or more independent variables on a dependent variable. The square of the multiple correlation coefficient, R^2, expresses the percentage of the variance in the dependent variable that is explained by the independent variable or variables. A corrected coefficient is also displayed; this is an unbiased estimate of the value of the coefficient in the population.

Multiple correlation coefficients are displayed only if there are three variables, or if there are four variables and the "2nd-order partials" option is selected. The corrected coefficient is computed only if the sample size is entered.

METHODS

If a value cannot be computed (e.g. because the coefficients on which a partial coefficient is based are incompatible), the program displays "?".

Partial correlation coefficients

Partial Pearson's correlation coefficients are computed by formulae 19.3 and 19.4 of Blalock (1979), and partial rank correlation coefficients by corresponding formulae (for Kendall's *tau*, see Siegel and Castellan 1988: 259, formula 9.13; for Spearman's *rho*, see Altman (1991: 296).

The significance of Pearson's partial correlation coefficients is tested by formulae 19.28 and 19.29 of Blalock (1979), and their 95% confidence intervals are estimated as

$$[\exp(2 * F) - 1] / [\exp(2 * F) + 1] \text{ to } [\exp(2 * G) - 1] / [\exp(2 * G) + 1]$$

where $F = Z - / \sqrt{(n - v - 1)}$

$G = Z - 1.96 / \sqrt{(n - v - 1)}$

v = total number of variables

$Z = \log((1 + r) / (1 - r)) * 0.5$

Confidence intervals for the difference between a simple coefficient and the corresponding first-order partial correlation coefficient are estimated by the procedure described as Model C by Olkin and Finn (1995), using their formulae 7 and 8 to compute the elements of the variance-covariance matrix.

The significance of first-degree *partial tau coefficients* is assessed by comparison with critical levels for one-tailed P = 0.05, 0.025, 0.01, 0.005, and 0.001 (Siegel and Castellan 1988: Table S) if the sample size is 20 or less. If the sample size exceeds 20 a large-sample Z test is used (Siegel and Castellan 1988: 260, formula 9.15).

If the sample size is 31 or less, the significance of first-degree *partial rho coefficients* is appraised by the use of critical levels for one-tailed P = 0.05, 0.025, 0.01, 0.005, and 0.001 (Siegel and Castellan 1988: Table Q), after

reducing the sample size by 1 (Altman 1991: 530); if the sample size exceeds 31 the following *t*-test is used (Altman 1991: 296):

t = sqrt{(N - 3) / [1 - (partial *rho*)^2 with (N - 3) degrees of freedom.

Multiple correlation coefficients

Multiple correlation coefficients are computed by formulae 19.20 and 19.21 of Blalock (1979) and their *significance* is tested by formula by Blalock's formula 19.27.

The *unbiased estimate of the population value* is estimated by Blalock's formula 19.24.

F5. SAMPLE SIZE AND POWER FOR TESTING A CORRELATION COEFFICIENT

This module computes the required *sample size* (the minimum number of subjects, i.e. of pairs of observations) and *power* for tests of the difference of a correlation coefficient from zero and (optionally) from a specified reference value.

The required significance level (*alpha*) must be entered, together with the required power (to compute the sample size) or sample size (to compute power). Optionally, the expected percentage of selected subjects expected to be lost because of refusal to participate or other reasons can also be entered.

The computed sample size is adjusted by inflating it (if necessary) to allow for losses (which of course does not compensate for possible selection bias), and then rounded up to the nearest whole number.

METHODS

The program uses formulae 19.18, 19.19 and 19.20 of Zar (1998).

The required sample size is rounded up to the nearest whole number, after making allowance (if necessary) for the percentage of expected losses (L%) by multiplying the number by $1 / [1 - (L / 100)]$.

F6. CALCULATION OF A CORRELATION COEFFICIENT FROM A PAIRED T-TEST RESULT

This module uses the result of a paired t-test to calculate a correlation coefficient between two variables.

It requires entry of the t value (or the two-tailed P value and the number of pairs of observations) and the two mean values and standard deviations.

It may help in the use of reports that provide a paired t-test but not a correlation coefficient.

METHODS

The formula, which is derived from equation 17.6 in Sheskin (2007), is:

$$r = [(SD_A / \sqrt{}(N))^2 + SD_B / \sqrt{}(N)^2 - (mean_A - mean_B)^2 / t^2] / (2 * SD_A / \sqrt{}(N) * SD_B / \sqrt{}(N))$$

where r = correlation coefficient
N = no. of pairs
$mean_A$ and $mean_B$ are the two means
SD_A and SD_B are the two standard deviations.

If t is not entered, it is derived from the P value and the degrees of freedom (N – 1).

F7. SAMPLE SIZE FOR ESTIMATION OF INTRACLASS CORRELATION COEFFICIENT

This module computes the sample size required for estimating an intraclass correlation coefficient with precision and assurance.

The intraclass correlation coefficient (ICC), which expresses the proportion of all variation that is not due to measurement error, is widely used in reliability studies that compare two or more observations of each subject, e.g. those comparing different observers, different methods, or different times. An ICC of <0.2 reflects "slight", 0.21 to 0.4 "fair", 0.41 to 0.6 "moderate", 0.61 to 0.8 "substantial", and above 0.8 "almost perfect" reliability, according to Landis and Koch (1977).

The module provides two methods of calculation. The first takes account of the desired lower level of the ICC's confidence interval, and the probability of achieving the desired precision. This method has been shown to be very accurate (Zou 2012). The second is based on the desired width of the confidence interval, which it assumes is symmetrical. Both methods require prespecification of the expected value of the ICC, and the desired probability (e.g. 80%) of reaching the required precision.

The procedure provided by this module is preferable to that provided by module S6 of PAIRSETC, which offers only a 50% chance of achieving the required precision.

METHODS

The method based on the desired lower level of the ICC's confidence interval uses formula 7 of Zou (2012), and the method based on the desired width of the confidence interval uses their formula 5.

G. ANALYSIS OF A CONTINGENCY TABLE LARGER THAN 2x2

This module analyzes a contingency table with 2-50 rows and 3-50 columns, providing measures of association and significance tests that appraise the association between two variables. The categories of the variables may be nominal or ordered. The module is not designed for comparisons of paired observations.

Several of the measures and tests are applicable to either nominal-scale or ordinal-scale variables. These are *Cramer's V, Sakoda's modified contingency coefficient, Cohen's effect-size index* (*w*), *Goodman and Kruskal's tau, Theil's uncertainty coefficient, odds ratios* (expressing the association of each row category with each column category), and conventional (Pearson) and log-likelihood-ratio (G2) *chi-square tests*, with *adjusted residuals*. It permits comparisons with a single selected row or column, performs *pairwise comparisons* of all rows and of all columns, and *allows chi-square to be partitioned* by combining (collapsing) categories. Haldane's large-table chi-square test is performed if there are 30 or more degrees of freedom. A *standardized version of the table* is provided.

If both variables have ordered categories, the relevant measures are Goodman and Kruskal's *gamma*, the *general odds ratio* and *general risk difference*, and *Spearman's and Kendall's rank correlation coefficient;* and a *chi-square test for trend* is performed. *Kruskal-Wallis one-way analysis of variance by ranks* is appropriate if only one variable has ordered categories.

Optionally, the module can examine *associations with multi-response variables* – it can analyse a table in which the categories of one or both of the variables are not mutually exclusive, i.e., where each subject may have entries in more than one category of the variable.

Optionally, the module can analyse a 2x3 table showing the results of a study with bilateral data, e.g. a randomized trial in which the outcome is reported in both eyes (or other paired parts of the body). *Donner's adjusted chi-square test* and *Rosner's tests* are performed, and 90%, 95% and 99% confidence limits are estimated for the difference between the two treatments.

Measures of association between categorical variables

Cramer's coefficient V (Siegel and Castellan 1988: 225-232) , *Sakoda's modified contingency coefficient, Goodman and Kruskal's tau, Theil's uncertainty coefficient*, and the *odds ratios* are measures of the strength of the association between two categorical variables. The categories may be nominal or ordinal, but their ordering does not affect these indices

Cramer's coefficient varies from 0 (no association) to 1 (complete dependence in a square table). It is based on chi-square and is regarded as a somewhat arbitrary measure; it gives greater weight to the columns or rows with the smallest marginal totals (Blalock 1979: 303-306). Its value (unlike that of chi-square) is not influenced by sample size .

Sakoda's contingency coefficient, a modification of Pearson's contingency coefficient, is also based on chi-square, and (unlike the Pearson coefficient) varies from 0 to 1. Like Cramer's coefficient, it can be interpreted as a proportion of the maximum variation between the variables.

Cohen's effect-size index (*w*) is computed from chi-square; it can exceed 1. By Cohen's criteria, 0.5 or more indicates a large effect size, 0.3 or more (but less than 0.5) indicates a medium effect size, and 0.1 or more (but less than 0.3) indicates a small effect size (Cohen 1988: 222 – 226). Cohen warns that these criteria should be used only when there is no better basis for evaluation. An adjusted *w*, controlling for the size of the table, is also computed, as suggested by Sheskin (2007: 658).

Goodman and Kruskal's tau expresses the extent to which knowledge of one of the variables enhances the accuracy with which the other can be predicted (Blalock 1979: 307-310; Jacobson 1976: 430-434; Agresti 1990: 24-25). It varies from 0, which means that the one variable is no help in predicting the other, to 1, which means that the one variable perfectly specifies the other. Goodman and Kruskal's *tau* is calculated for predictions in each direction; a symmetric (nondirectional) version is also computed. *Tau* tends to become smaller as the number of categories increases.

Theil's uncertainty coefficient is another measure of the extent to which knowledge of one of the variables enhances the accuracy with which the other can be predicted. It varies from 0, which means that the one variable is no help in predicting the other, to 1, which means that the one variable perfectly specifies the other. The coefficient is calculated for predictions in each direction; a symmetric (nondirectional) version is also computed.

The *odds ratios* that are displayed express the associations between each row category and each column category. They should be treated with caution, as their confidence intervals may be wide unless numbers are large. An odds ratio above 1 indicates a positive association. If the table has more than 100 cells, the odds ratios are displayed only if "Show very detailed results" is checked.

Chi-square tests

Pearson (conventional) and *log-likelihood-ratio chi-square tests* generally lead to the same conclusions. When they do not, many statisticians prefer the log-likelihood-ratio test (Zar 1996: 503). If Williams's criterion for preferring the log-likelihood-ratio chi-square to the Pearson chi-square is met – i.e. if any expected frequency (under the null hypothesis) is less than its difference from the observed frequency (Williams 1976) – the program displays a message to this effect.

Chi-square tests may be misleading if the expected frequencies (under the null hypothesis) are too small. Cochran (1954) recommended that fewer than one-fifth of the cells should have expected frequencies of less than 5, and none should have an expected frequency of less than 1. The program displays a warning if these conditions are not met. A warning is also shown if the mean frequency per cell is under 5, since the likelihood-ratio test may then be of low validity; the P-value tends to be too high if most expected values are less than 0.5, and too low if most expected values are between 0.5 and 5 (Agresti 1996: 194).

Haldane's large-table chi-square test (Maxwell 1961: 41-44) is performed if there are 30 or more degrees of freedom. This test is based on the exact mean and variance of chi-square (Maxwell 1961: 41-44), and its validity is not affected by zeroes or small cell frequencies. Two alternative P values are displayed, based on Dawson's and Bartlett's modifications respectively.

Comparisons of rows or columns

The program performs *pairwise comparisons* of all rows and of all columns, using likelihood-ratio chi-square tests, and providing two P values in each instance – one appropriate for a planned test of an *a priori* hypothesis, and one applying Sidak and Bonferroni adjustments in order to compensate for multiple testing.

The Sidak and Bonferroni adjustments both assume that the comparisons are independent. The Sidak adjustment is slightly less "pessimistic" (Abdi 2007) - i.e., less severe, less conservative, and it has a bit more power than the Bonferroni method. So from a purely conceptual point of view, the Šídák method may be preferred). If the assumption of independence is false, both procedures "do a good job of protecting against false statements of statistical significance, but have less power to detect real differences" (GraphPad Statistics Guide 2013).

The program also permits comparisons with a single selected reference row or column, providing Sidak and Bonferroni-adjusted P values.

Adjusted residuals

Adjusted residuals, which show which cells contribute most to the chi-square, may be helpful in determining the sources of a significant association. The residuals are the discrepancies between the observed frequencies and the values expected under the null hypothesis, converted to *Z* scores so as to indicate their statistical significance. An adjusted residual over 1.96 or under -1.96 indicates significance at the $P < 0.05$ level, and an adjusted residual over 2.58 or under -2.58 indicates significance at the $P < 0.01$ level. The use of this procedure is described by Everitt (1977: 46-48) and Agresti (1996:31-32).

If the table has more than 100 cells, the adjusted residuals are displayed only if "Show very detailed results" is checked.

Partitioning of chi-square

Options are offered for comparisons of each row with each other row, and of each column with each other column. These may be useful if one of the categories is a reference or control group. The P values are adjusted for multiple comparisons.

Options are also offered for the combination (collapsing) of selected rows, selected columns, or selected rows and columns. The selected rows or columns need not be adjacent ones. For explanations of some of the possibilities, see Armitage and Berry (2002: 514-516) or Siegel and Castellan (1988 194-198). Two sets of P values are displayed - one suitable for the testing of *a priori* hypotheses, and one for safe use even if hypotheses were suggested by the data.

Associations between ordinal variables

The following measures are appropriate if both the row variable and the column variable have categories that fall into a natural order.

Two *coefficients of rank correlation* are provided, namely Spearman's *rho* and Kendall's *tau b*, These have different numerical values but are similar in their ability to appraise the significance of associations (Siegel and Castellan 1988: 251). One-tailed and two-tailed P values are displayed. The significance of *tau b* is tested by a large-sample method, and P should be regarded as approximate if the sample is small.

Goodman and Kruskal's *gamma*, which ranges from -1 to 1, expresses the difference between the probability that, in a randomly selected pair of observations, a higher value of one variable is accompanied by a higher value of the other variable (concordance) and the probability that a higher value of one variable is accompanied by a lower value of the other variable (discordance), when tied observations are ignored). Confidence intervals (90%, ,95%and 99%) are reported.

The *general odds ratio* (Edwardes and Baltzan 2000), which is computed from gamma, is an estimate of the ratio of concordant to discordant pairs of observations; it is Agresti's *alpha* (Agresti 1980). If the variables represent exposure to a risk or protective factor, and a disease or other outcome, the general odds ratio expresses "a type of shift of median severity as exposure increases", but is not affected by the distances between severity categories (Edwardes and Baltzan 2000). It is applicable at least to cross-sectional studies, unmatched case-control studies, cohort studies comparing different exposure categories, and two-armed randomized control trials. Confidence intervals (90%, 95%, and 99% are reported.

The *general risk difference* (Edwardes and Baltzan 2000), which is *Somers' d*, is a weighted average of the risk differences seen in the component 2 x 2 tables that can be constructed from the large *(r* x *c*) table. Two alternative values are reported, their applicability depending on which of the two variables is the outcome variable. The measure is applicable at least to cross-sectional studies, cohort studies comparing different exposure categories, and two-armed randomized control trials.

Kruskal-Wallis one-way analysis of variance by ranks

This analysis (Siegel and Castellan 1988: 206-216; Sprent 1993: 138-141, 226-228) is appropriate if one variable has ordered categories and the other has not. The analysis is done twice. The first analysis is appropriate if the column variable has ordered categories; it tests the null hypothesis that the distribution in the ordered column categories is the same in all row categories. The second analysis is appropriate if the row variable has ordered categories ף.it tests the null hypothesis that the distribution in the ordered row categories is the same in all column categories.

A large-sample approximation is used, treating the Kruskal-Wallis statistic as chi-square; the result should be treated with reserve if the samples are very small. The P values may be regarded as two-tailed.

Test for trend

A chi-square test for trend (the "Mantel-Haenszel chi-square), based on scores (1, 2, 3, etc.) allocated to the categories, is appropriate if both variables have ordered categories (Armitage and Berry 2002: 509-511). The overall chi-square is partitioned into two components, one expressing the effect of the linear regression, and one expressing departure from linear regression.

Associations with multi-response variables

Optionally, the module can analyse a table in which the categories of one or both of the variables are not mutually exclusive, i.e., where each subject may have entries in more than one category of the variable. The table might, for example, compare the symptoms of different groups of subjects, where each subject may have more than one symptom, or it might show responses to a multiple-response ("pick any of the following") survey question, or to two multiple-response questions.

It provides two alternative summary chi-square tests for marginal independence between a single-response variable (whose mutually exclusive categories are entered in separate rows) and a multi-response variable (whose categories are entered in separate columns), or between two multi-response variables. The tests use the sum of the chi-square values and degrees of freedom for the associations in separate components of the table.

The first summary chi-square test is based on the associations between the "row" variable (which may be a single-response or multi-response one) and each category of the "column" (multi-response) variable. An *r* x 2 table (where *r* is the number of categories in the "row" variable) is constructed for each category of the "column" variable, showing the association between the row variable and one category of the column variable, and the chi-squares and degrees of freedom in the various tables are summed. This" naive" summary chi-square statistic (Agresti and Liu 1999) can be regarded as a first-order member of the Rao-Scott family of tests (DFecady and Thomas 2004). It is an approximate test, and may be "liberal" (giving an unduly low P value) if there are large inter-item correlations .

The second summary chi-square test is similar, but is based on a set of 2 x 2 tables constructed to show the association between each category of the "row" variable and each category of the "column" variable; the chi-squares and degrees of freedom in the various tables are summed (Vlach and Plasil, undated; Bilder and Loughlin 2004). This test too may be "liberal"if there are large inter-item correlations.

The chi-squares in the separate component tables (*r* x 2 and 2 x 2 tables) are reported, together with P values that have been adjusted by the Bonferroni method to compensate for multiple testing. Each of the summary chi-square tests is accompanied by an overall test that uses the lowest of its component Bonferroni-adjusted P values as an overall test of multiple marginal independence, a "valid albeit somewhat conservative way of simultaneously using the ... marginal Pearson statistics to test multiple independence ... When [the] overall test gives evidence against the null hypothesis, the separate chi-squared components provide information about the marginal tables that are responsible" (Agresti and Liu 1999). The Bonferroni-adjusted tests are likely to be especially conservative if the variables have many categories (Bilder and Loughlin 2004).

The odds ratios in the 2x2 tables are reported as well as the chi-squares, and these too may throw light on the overall finding.

Studies with bilateral data

This analysis is applicable to a study with bilateral data, e.g. a trial in which randomly selected subjects receive different treatments, and the occurrence of a specified outcome is reported in both eyes (or other paired parts of the body). The data required, for each treatment, are the numbers of subjects with the specified outcome on neither side, on one side, or on both sides. A treatment may be compared with another treatment, with a control procedure, or with no treatment.

The analysis takes account of the probable correlation between the occurrence of the specified outcome in the two eyes [etc.] of the same subject.

The tests performed are *Donner's adjusted chi-square tes*t, which uses an empirical estimate of the intraclass correlation between the responses in the two eyes of the same person, and provides a P value considerably higherjr than that of an unadjusted chi-square test that ignores this correlation (Donner 1989), and two t*ests proposed by Rosner* (1982), one assuming complete independence between the findings on the two sides, and one assuming that the outcome in the two eyes of the same subject are dependent.

The program estimates 90%, 95% and 99% confidence intervals for the difference between the proportions of eyes with the specified outcome in the two treatment groups, using methods based on Wald-type statistics (Tang et al. 2011). Two sets of intervals are reported, based respectively on dependence and independence models.

All these procedures have been validated by computer simulation studies.

Table standardization

Table standardization (a form of "*raking*") may facilitate the comparison of similar tables that have different row and/or column totals, e.g. tables referring to different populations or different times.

The method used is iterative proportional fitting (IPF), which adjusts the values in the cells so that they add up to selected (standard) marginal totals. ETCETERA's procedure is based on equal sizes for the "row" marginals and equal sizes for the "column" marginals. The adjusted values are displayed as percentages of the table's grand total.

METHODS

Measures of association between categorical variables

Cramer's coefficient V is calculated from chi-square (Siegel and Castellan 1988: formula 9.1).

The formula for *Sakoda's modified contingency coefficient* is

$$C / \sqrt{[(k - 1) / k]}$$

where C = Pearson's contingency coefficient

$$= \sqrt{[\text{chi-square} / (\text{chi-square} + N)]}$$

N = total number of observations

k = number of columns or number of rows, whichever is smaller.

Cohen's effect-size index (w) is computed by the formula

$$w = \sqrt{(\text{chi-square} / N)}$$ (Volker 2006: formula 17).

The adjusted w takes account of the size of the table by using Sakoda's contingency coefficient S:

$$w = \sqrt{(S^2 / (1 - S^2))}$$ (Sheskin 2007: 658)

The *odds ratios* expressing the associations between each row category and each column category.are computed by collapsing the table to a 2x2 table for each pair of pair of categories.

Goodman and Kruskal's tau (Agresti 1990: 24) is computed twice, with fixed marginal totals for the row and column variables in turn; a symmetric version is also computed. For detailed formulae, see Jacobson 1976.

A convenient formulation of the asymmetric and symmetric versions of *Theil's uncertainty coefficient* is available on the Internet at http://www.statisticssolutions.com/Nominal-Association.htm.

Chi-square tests

Formulae for chi-square are provided by most statistics textbooks (e.g. Zar (1998: formula 23.1 for Pearson's chi-square and 23.11 for the likelihood ratio test). The computation of likelihood-ratio chi-squares when there is a zero frequency is made possible by changing the zero to 0.0000001; an appropriate message is displayed.

Formulae for the computation and appraisal of *Haldane's large-table chi-square test* are provided by Maxwell (1961: 41-44). Expressions provided by Dawson (formula 2.3) and Bartlett (formula 2.5) are used.

Comparisons of rows or columns

To compensate for multiple comparisons, the P value is multiplied by the number of comparisons, i.e. by $a / (a - 1) / 2$ when all pairs of rows or columns are compared, and by $a - 1$ when comparisons are made with a single row or category,
where a = number of rows or categories.

Adjusted residuals

See Haberman (1973), Everitt (1977: formulae 3.6 to 3.8) or Agresti 1996: formula 2.4.4).

Partitioning of chi-square

See Armitage and Berry (2002: 516) or Siegel and Castellan (1988 194-198).8(

Kruskal-Wallis one-way analysis of variance by ranks

Formulae for the Kruskal-Wallis test are provided by Siegel and Castellan (1988). The Kruskal-Wallis statistic is corrected for ties (formula 8.5, p 210).

Test for trend

The test for trend is described by Armitage and Berry (2002: 509-511). Formula 15.12 is used. The overall chi-square is partitioned as described by Maxwell 1961: 71.

Associations between ordinal variables

Spearman's *rho* is computed by a formula that takes account of tied ranks (Siegel and Castellan 1988: 241, formula 9.7). If there are 30 or fewer observations, the significance of *rho* is appraised by the use of critical levels for one-tailed P = 0.05, 0.05, 0.01, 0.005, and 0.001 (Siegel and Castellan 1988: Table Q). If N > 30, a *t*-test is used (Siegel and Castellan 1988: 243, footnote), based on the null variance.

Kendall's *tau b* is calculated by a formula that makes allowance for tied observations (Siegel and Castellan 1988: 249, formula 9.10). The program uses the *kend12* algorithm of Press *et al*. (1989: 542-543).

The computation of Goodman and Kruskal's *gamma* and Somers' *d* (which is reported as the *general risk difference*) is described by (*inter alii*) Siegel and Castellan (1988: 291-298 and 303-310).

Confidence intervals for *gamma* are estimated by estimator 9 of Lui and Cumberland (2004), as recommended on the basis of their computer simulations. In accordance with their recommendation, if any cell in the table has a zero value, 0.5 is first added to all cells.

The *general odds ratio* is computed as (1 + *gamma*) / (1 - *gamma*), as proposed by Edwardes and Baltzan (2000). Its confidence limits are derived similarly, from the confidence limits of *gamma*.

Associations with multi-response variables

The summary chi-square and overall (Bonferroni-adjusted) tests are described by Agresti and Liu (1999), and Vlach and Pasil (undated: formula 4). The P values are Bonferroni-adjusted by multiplying them by *c* (for *r* x 2 tables) or by *rc* (for 2 x 2 tables),
where *r* = number of categories in the "row" variable
c = number of categories in the "column" variable.

Sidak adjusted P = 1 - (1 - unadjusted P)^*K*, where *K* = no. of tests..

If there is a zero cell in any of the component 2x2 tables, 0.5 is added to each cell in the table.

Studies with bilateral data

No adjustment is made to the observed cell totals.

The computation of Donner's adjusted chi-square is explained by Donner (1989: 607-608).

Rosner's test statistics (TRD and TRI) are computed by the formulae provided by Tang et al. (2008: 3723-3724), and are evaluated by the aymptotic test method. The measure of dependence used for this purpose (R) is estimated by a formula provided by Rosner (1982: 109).

Formulae for the confidence intervals for the difference between proportions (based on Wald-type statistics) are provided by Tang et al. (2011: 236). No adjustment is made to the cell totals.

Table standardization

Each row of values is proportionally adjusted to conform with the desired row margins (equal sizes for the "row" categories), and then each column of row-adjusted values is proportionally adjusted to conform with the desired column margins (equal sizes for the "column" categories). These steps are repeated until further reiteration makes only a negligible difference (less than 0.00001).

The procedure is explained in detail (using Excel) by Charles Zaiontz (2015).t

G2 RAKING

This module renders the findings in a sample more representative of the population, in instances where the sampling ratio varies in different categories of subjects (because of different response rates or for other reasons). It is applicable to a contingency table (of 2-50 .rows and up to 9 columns) that shows the relationships between two sets of categorie

The population data ("census data", "control data") for each category must be entered as well .as the table.

The raked data (after 100 iterations) is reported, together with *suggested weights* for use in analyses of the sample data. The module also reports the *design effect, the design factor,* and the *effective sample size* (N^) resulting from weighting.
.

Raking (sample balancing)

Each row and each column in the table may refer to a single characteristic (e.g. male, or female, or ”40+ years of age" or “<40 years of age”) or (if the appropriate census data are available) to a combination of characteristics (e.g. males over 40 years of age, etc.).

The raking procedure is stepwise. It has two stages, which the program then repeats 100 times. In the first stage a modification is made to the numbers in each row, based on the relationship between the row total and the corresponding census total. This alters the column totals. In the second stage a modification is made to the numbers in each column, based on the relationship between the column total and the corresponding census total. This alters the row totals. These two steps are then repeated again and again, and convergence (ie., almost no further changes in the row or column totals) is ultimately reached.

This process neutralizes the effect of different sampling ratios in the various categories, but it does not handle any other biases. It assumes that the sample in each row or column is representative of the population in that category.

Design effect and design factor

The *design effect* and the *design factor* are expressions of the reduction in the precision of estimates as a result of weighting. The design effect is the ratio of the variance after weighting to the variance in a simple random sample, and the design factor is the factor by which the standard error (and hence the confidence interval) is multiplied by the weighting.

Effective sample size

The *effective sample size* (N^) estimates the number of subjects required by a study using simple random sampling in order to yield the same sampling error as a study using the weighted data. It provides an indication of the loss of power because of weighting.

METHODS

Raking

The basic formula used in raking is, for each cell in each row [or column]:

weighted value = value x N / n

where value = each value in the row [or column]
N = census total
n = sample total for the row [or column]
(Battaglia *et a*l. (2009)

Design effect and design factor

Design effect = Variance after weighting / variance for a simple random sample
Design factor = √(design effect)

Effective sample size

Effective sample size (N^) = sample size / design effect

H. MEDIAN POLISH OR MEAN POLISH OF A TWO-WAY TABLE

This module applies the median polish or mean polish procedure to a two-way table (with up to 50 columns and up to 200 rows). It fits a *model representing the additive or multiplicative effects* of the row and column variables, reports the *deviations from the mode*l., and displays the *pattern of the deviations*. It appraises *goodness of fit* and computes the residual (unexplained) variation, which may point to statistical interaction, e.g. to a cohort effect if the variables are time and age.

The values in the table may be numbers of any kind – frequencies, measurements, proportions, or rates.

Median polish differs from mean polish in that the analysis uses medians and not means, giving less weight to extreme values.

Median polish

The median polish procedure fits an additive or multiplicative model, representing the additive or multiplicative effects of the row and column variables, to a two-way table. This is done by subtracting the row median from each value, then subtracting the column median, and repeating these two steps until they produce no further change.

Multiplicative effects are appraised by using the logs of the values shown in the table.

The row and column effects (respectively) are reported in terms of the differences (in the additive or multiplicative model) from (respectively) row 1 and column1, which can be used as reference categories.

The values in the table may be numbers of any kind – frequencies, measurements, proportions, or rates. The values in each row must be separated by spaces. Since the rows in the table have a limited available length, difficulty may be encountered if there are many columns; it may be necessary to reduce the number of decimal places in order to ensure that the values are spaced.

The procedure and its epidemiological applications are described by Selvin (2004: 100-110).

Mean polish

Mean polish is performed in the same way, but using the row and column means instead of their medians. This gives more weight to extreme values, and is less robust than median polish.

Goodness of fit

The program reports the proportion of the total variation that is accounted for by the combined effects (additive or multiplicative) of the row and column variables, and the residual proportion that is not explained by these effects. The unexplained variation may point to statistical interaction, e.g. to a cohort effect if the variables are time and age.

Deviations from the model

The program reports the deviations of the observed data from the adjusted values in the model, in terms of arithmetical differences (if the model is additive) or ratios (if the model is multiplicative).

To facilitate detection of patterns, the deviations are displayed as symbols as well as numerically.
For the additive model, the symbols are ++, +, o, -, and --.
For the multiplicative model, they are +++, ++, +, =,-, --, and ---.

METHODS

Median polish and mean polish

See Selvin (2004: 100-110).

Goodness of fit

The proportion of variation accounted for by the row and column effects (Emerson and Wong1985; cited by Amali *et al.* (1997) is

$$1 - \Sigma\, r_{ij} \,/\, \Sigma\,(y_{ij} - M)$$

and the unexplained variation is $\Sigma\, r_{ij} \,/\, \Sigma\,(y_{ij} - M)$

where r_{ij} = residual value in row i and column j
y_{ij} = observed value in row i and column j
M = overall median or mean

I. ANALYSIS OF A THREE-WAY CONTINGENCY TABLE (LOGLINEAR ANALYSIS)

This module analyzes a three-way contingency table in which each of the three variables has two to four categories. It is not necessary to specify a dependent variable.

The module fits a number of *loglinear models* to the observed frequencies and evaluates and compares their *goodness-of-fit*, to permit appraisal of the relative importance of different effects.

If there are binary (two-category) variables, it provides *odds ratios* that express their association.

Loglinear models

Loglinear analysis appraises association and interaction patterns among a set of categorical variables. Its application to three-way tables is explained in detail by (inter alia) Agresti (1996: 150-162 and 1990: 135-150).

This program performs a limited loglinear analysis. It uses the following loglinear models for the relationships between variables *A*, *B*, and *C*:

Models *AB*, *AC*, and *BC*, which represent two-way associations, in each case ignoring the third variable.

Model *A,B,C*, which expresses complete independence of the three variables.

Models *AB,C*, *AC,B*, and *BC,A*, which express partial independence.
In model *AB,C*, variables *A* and *B* are jointly independent of *C* – variables *A* and *B* may or may not be related, but neither is related to *C*; and variable *C* is independent of *A* and *B*.
In model *AC,B*, variables *A* and *C* are jointly independent of *B* – variables *A* and *C* may or may not be related, but neither is related to *B*; and variable *B* is independent of *A* and *C*.
In model *BC,A*, variables *B* and *C* are jointly independent of *A* – variables *B* and *C* may or may not be related, but neither is related to *A*; and variable *A* is independent of *B* and *C*..

Models *AC,BC*, *AB,BC*, and *AB,AC*, which express conditional independence:
Model *AC,BC* expresses the relationship between variables *A* and *B* when *C* is controlled; if a relationship is found between *A* and *B*, this might be explained by *C*.
Model *AB,BC* expresses the relationship between variables *A* and *C* when *B* is controlled.
Model *AB,AC* expresses the relationship between variables *B* and *C* when *A* is controlled.

In addition, the conditional independence of each pair of variables in the separate categories of the third variable is examined.

Goodness-of-fit tests

Log-linear chi-square tests are used to appraise the goodness of fit of the models. The program reports the chi-square, with its degrees of freedom, and the associated P value, and specifies the null hypothesis, which is that the variables are not related..

If the P value is under 0.05, the null hypothesis (of independence) is rejected. If P > 0.1, this is taken to indicate an adequate fit, and "good fit" is reported.

The results may indicate that a variable or association can be ignored, or that it must be taken into account because of its modifying or possibly confounding effect.

The fit of different models can be compared by taking the difference between their goodness-of-fit chi-squares and determining the relevant P value (using the difference between the degrees of freedom of the two tests). The program provides these comparisons of models with 'good fits'. A nonsignificant result indicates that the two models do not differ significantly in their goodness-of-fit, and the more parsimonious model, i.e. the one based on less information, may be preferred.

The goodness-of-fit results may be misleading if data are sparse. A warning is displayed if the total sample size is less than the recommended minimum, which is five times the number of cells in the three-way table.

Odds ratios

If there are at least two binary (two-category) variables, odds ratios (with their approximate 95% confidence intervals) are displayed to express their association, both when the third variable is ignored or controlled, and for each separate category of the third variable. The odds ratio when the third variable is controlled is computed by the Mantel-Haenszel procedure (apparent inconsistencies may be due to the fact that this procedure uses the raw data, whereas 0.5 is added to each cell frequency before calculation of the other odds ratios).

METHODS

Odds ratios

Odds ratios are computed after adding 0.5 to each cell frequency in the relevant 2x2 table (Fleiss *et al.* 2003, formula 6.20).

An approximate 95% confidence interval for the odds ratio (OR) is estimated by the formulae

$$\exp[\ln(OR) - 1.96(se)] \text{ and}$$
$$\exp[\ln(OR) + 1.96(se)]$$

where *se*, the standard error of ln(OR), is calculated from the cell frequencies *a*, *b*, *c*, and *d*, by formula 6.33 of Fleiss *et al.* 2003:

$$se = \sqrt{[1 / (a + 0.5) + 1 / (b + 0.5) + 1 / (c + 0.5) + 1 / (d + 0.5)]}$$

The Mantel-Haenszel odds ratio is computed by formula 10.52 of Fleiss (2003), and the estimation of its confidence intervals is described by Robins, Breslow and Greenland (1986) and by Rothman (1986: 219-220).

Goodness-of-fit tests

Log-linear chi-square tests are used, after converting any zero values to 0.0000001.

For the total three-way table (log-linear model *ABC*), use if made of formula 23.24 of Zar 1998.

For the component two-way tables, ignoring the third variable (models *AB*, *AC*, and *BC*) and in separate categories of the third variable, Zar's formula 23.11 is used (with 1 degree of freedom).

Chi-square values for the other models are derived from the above chi-squares by subtraction:
The chi-square for the *AB,C* model is the difference between the chi-squares for the *A,B,C* and *AB* models.
The chi-square for the *AC,B* model is the difference between the chi-squares for the *A,B,C* and *AC* models.
The chi-square for the *BC,A* model is the difference between the chi-squares for the *A,B,C* and *BC* models.
The chi-square for the *AB,BC* model is the difference between the chi-square for the *A,B,C* model and the sum of the chi-squares for the *A**B*** and *BC* models.
The chi-square for the *AC,BC* model is the difference between the chi-square for the *A,B,C* model and the sum of the chi-squares for the *AC* and *BC* models.
The chi-square for the *AB,AC* model is the difference between the chi-square for the *A,B,C* model and the sum of the chi-squares for the *AB* and *AC* models.

The degrees of freedom for all the tests are listed in Table 6.5 of Agresti (1990).

J. REGRESSION

This module performs linear regression for a model with up to seven independent variables, appraising their additive effects on a dependent variable. It computes a *regression equation*, based on least-squares regression analysis, and tests the significance of coefficients. It reports the *coefficient of determination* (*R*-squared), the *adjusted coefficient of determination,* Cohen's *f-squared*, and the *standard error of the estima*te, and provides an *analysis of variance* and an F-test. Outliers are reported.

If there is a single independent variable, *simple linear regression* is performed and the distribution of *residuals* is displayed, and additional regression analyses (including *exponential, elasticity*, and *nonparametric regression*) are performed. *Correlation coefficients* are computed, including (for multiple regression) partial correlation coefficients between the dependent variable and each predictor. Options are offered for *interrupted time-series regression*, and for the *comparison of regression coefficients*.

If multiple regression is performed, an *interaction* term or two interaction terms can be included in the model, and analyses are done with and without interactions; an F-test compares the two *R*-squared values. A backward elimination (step-down) option is provided (if there are no interaction terms), permitting the removal of one chosen independent variable at a time. A partial *F*-test assesses the significance of the change in *R*-squared.

If multiple regression is performed, the *level-importance* of each independent variable is reported.

The results may be unreliable if the sample size is small. The module provides estimates of the *sample size* required to attain a power of.80 (with alpha = .05). This facility can be used to estimate the sample size required for any regression analysis.

The module uses *G-computation* (based on the multiple regression coefficients) to estimate the effect of a dichotomous variable that is involved in an interaction with another variable or variables.

Regression equations

The regression equations, which comprise intercept and slope coefficient for each of the independent variables, is computed by the ordinary least squares method. Two-tailed P values are provided for the coefficients. An F-test appraises the significance of the model, the null hypothesis being that there is no relationship between the independent and dependent variables. The regression equation, which comprises a constant and a coefficient for each of the independent variables, is computed by the ordinary least squares method. Two-tailed P values are provided for the coefficients. An *F*-test appraises the significance of the model, the null hypothesis being that there is no relationship between the independent and dependent variables.

The standard error of the estimate is provided, as an indication of the accuracy of predictions that use the regression equation. It is the standard deviation of the residuals. For large samples, the standard error of the estimate approximates the standard error of a predicted value.

If there is one independent variable, the regression of the dependent variable on the independent variable is supplemented by the regression of the log10-transformed dependent variable on the independent variable (unless this is prevented by zero or negative values), *exponentia*l *regression*, and *constant elasticity regression*. The regression lines are shown in graphs (see below), together with a graph showing the distribution of the deviations from the simple regression. If the deviations are equally distributed above and below zero, this is evidence of homoscedasticity (equality of variation) which is an assumption of regression analysis. If the distances from zero tend to increase as the value of the independent variable increases, possibly creating a fan-like or cone-like appearance, this is evidence of heteroscedasticity, and may justify the use of log transformation. The *elasticity* curve shows the percentage rise in the *Y* variable that is associated with the same [percentage rise in the *X* variable.

The *nonparametric procedure*, which does not assume a normal distribution, has the advantage of robustness – i.e., discrepant 'outlier' observations have a reduced effect; two estimators of the intercept may be shown; the second is recommended if deviations from the regression line can be assumed to be symmetrical.

Interaction terms

An interaction term (e.g. “height*age”) expresses the joint effect of two of the independent variables, each of which modifies the effect of the other. The product of the values of the two variables is treated as an additional term in the regression model. An interaction expresses a multiplicative relationship, and if present it indicates a departure from simple additivity.

The inclusion of interaction terms in the model is optional. Up to six independent variables and one interaction term can be entered, or up to five independent variables and two interaction terms.

An *F*-test compares the *R*-squared values obtained when the model includes or excludes the interaction(s).

Correlation coefficients

In simple regression analysis the program computes the correlation coefficient, and in multiple regression analysis it computes correlation coefficients between the dependent variable and each predictor – both the simple bivariate zero-order coefficients and the partial correlation coefficients (controlling for all other predictors). The significance of each coefficient in comparison with zero is computed, and the corresponding coefficients of determination (*R*-squared) or partial determination (r-squared) are displayed.

Coefficients of determination

R-squared (the coefficient of determination) can be interpreted as the proportion of variation in the dependent variable that is explained by the independent variables. It is not a satisfactory measure of the goodness of fit of the regression model.

The adjusted coefficient of determination is an acceptable measure of the goodness of fit of the regression model, and is a better estimate than *R*-squared of the population coefficient of determination. It incorporates a downward adjustment to compensate for the possible effect of the number of independent variables on the residual variance. It may be negative if the population coefficient is near zero (Zar 1998: 423). It is displayed in multiple regression analyses.

Cohen's *f-squared*

Cohen's f-squared may be used as a measure of effect size. It is computed from *R*-squared, and can exceed 1. By Cohen's criteria, 0.35 or more indicates a large effect size [equivalent to aR-squared value of 0.51), 0.15 or more (but less than 0.35) indicates a medium effect size, and 0.02 or more (but less than 0.15) indicates a small effect size (Cohen 1988: 410-414). Cohen warns that these criteria (based on social science research) should be used only when there is no better basis for evaluation.

Effect of removing a variable

If a variable is removed from the model, the program reports the change in *R*-squared (the marginal R-squared), and performs an F test that assesses the significance of the change.

Outliers

The program displays a list of outliers (if any), i.e. cases where the prediction based on the regression equation is very far from the observed value of the dependent variable.

Sample size

If the sample is small, tests may be insufficiently powerful and the results may be unreliable. The module provides estimates of the sample size required to attain a power of .80 (with *alpha* = .05), for a regression analysis and for testing partial correlation coefficients, for comparison with the actual sample size. These estimates are based only on the number of predictors and the strength of the association, as reflected by coefficients of determination.

Estimates of the sample size required for a regression analysis are provided for selected coefficients of determination ranging from 0.02, indicative of a weak association (i.e., a correlation coefficient of 0.1), to 0.26 (i.e., a strong association, with a correlation coefficient of 0.51). Estimates of the sample size required for testing partial correlation coefficients are provided for selected coefficients of determination between 0.01 (weak) and 0.26 (strong), and also for the partial correlation coefficients reported for the observed data.

This module may also be used to estimate the sample size required for any regression analysis with up to seven predictors or for computing a partial correlation coefficient, by entering the number of predictors (or, for a partial correlation coefficient, the number of variables held constant plus one) and then by entering imaginary data and (ignoring all the results except those concerning sample size, which are reported at the end of the output) finding the sample size corresponding to the expected coefficient of determination or partial determination in the proposed study.

The estimates of sample size are based on a rule-of-thumb method suggested by Harris (1975), as modified by Green (1991), a new rule-of-thumb method suggested by Green (1991), and a newer method proposed by Maxwell (2000). These methods are fairly accurate in comparison with power analyses if there are fewer than seven predictors, and then become more conservative. If the association is strong, they tend to overestimate the sample size if the association is weak and to underestimate it slightly if the association is strong, although the degree of underestimation is not great when there are few predictors (Green 1991). The discrepancies from power analyses are slight if the strength of the association is medium, or the number of predictors is small.

Maxwell *et al.* (2008) point out that these sample sizes may be appropriate if the purpose of the study is to appraise the significance of findings, but may often underestimate or (sometimes) overestimate the sample size required to provide precise estimates of parameters (i.e., with narrow confidence intervals).

Level-importance

This statistic (Achen 1982) expresses the influence in this sample of each independent variable on the level of the dependent variable. Assuming causality, it is the net change in the dependent variable's level attributable to each independent variable. This is akin to the elasticity concept commonly used in economics, expressing the percent change in a dependent variable for a 1% change in an independent variable (Kruskal and Majors 1989). Since the sum of the level-importance statistics (plus the intercept) is precisely the mean of the dependent variable, the level-importance of each variable can also be expressed as a percentage of the mean of the dependent variable.

Exponential regression

If a single independent variable is entered, exponential regression is performed, and the formula for the best-fit exponential curve is computed and graphed. This computation is omitted if any value of the independent variable is zero or less.

The *coefficient of determination* (*r*-squared) is computed. This expresses the proportion of the variation in the log of the dependent variable that can be explained by the relationship with the independent variable..

G-computation

If one of the variables is a dichotomous (Yes/No) variable (coded 1/0) whose effect is modified by another variable or variables, a single estimate of this effect (the marginal causal treatment effect) is calculated by G-computation (Snowden et al. 2011). This requires entry of the binary valuable as the second in the list of variables, and the inclusion in the regression model of an interaction or two interactions with this variable, as well as suspected confounders. The model must include main terms for the modifiers .

The analysis is meaningful if the variable precedes the dependent (outcome) variable in time, and refers to a point-treatment (not time-varying) exposure. The procedure has been validated by computer simulation (Snowden *et al.* 2011). The estimated effect is equivalent to standardization using the distribution of covariates in the study sample as the standard Vansteelandt and Keiding (2011). "Application of this method", say Snowden *et al*. ,"allows investigators to use observational data to estimate parameters that would be obtained in a perfectly randomized controlled trial".

If interaction is the primary concern, for example in clinical settings where the effectivity of treatment varies in different groups, the conditional estimates of effect that are provided by regression analysis are of course of more interest than the estimate provided by G-computation.

Interrupted time-series regression *(ITS)*

The program compares the trend of a variable before and after the occurrence of a public health intervention or other event, such as the banning of smoking in public places (Bernal *et al*. 2016), or an economic crisis or the outbreak of a war. Simple linear regression analyses are performed for both periods, with tests for autocorrelation (serial correlation) for both periods; if either of these tests suggests that there is significant autocorrelation, the Cochrane-Orcutt procedure is used to adjust the regression coefficients. The difference between the "before" and "at or after" slopes (the *b* coefficients) is tested, using the adjusted coefficients if they are available. Welch's *t*-test, which is appropriate even when variances are unequal, is used for this purpose. The mean values in the two periods are also compared, using Welch's *t*-test. The difference between the counterfactual and actual lines at the end of the study is reported.

In studies of time trends, *autocorrelation of residuals* (correlation between the deviations of consecutive values from the regression line) may be caused by factors that have an effect persisting over successive periods, and that do not find their expression in the straight regression line; they may or may not be confounders of the association under study, such as fluctuations in diagnostic criteria. Auto correlation will produce an unduly narrow confidence interval for the slope coefficient, and its presence may throw doubt on the appropriateness of a straight regression line. Two tests for autocorrelation are performed - *the Durbin-Watson test*, which assumes a normal distribution for the residuals, and *a runs test*, which makes no such assumption. Two-tailed and one-tailed *P* values (testing for positive and negative correlation) are displayed; a low *P* value indicates auto correlation. If the runs test or the Durbin-Watson test suggests significant autocorrelation, the *Cochrane-Orcutt procedure* is used to produce adjusted

regression coefficients for that subset of data (the "before" or "at or after" data; if available, the adjusted coefficients are used in the comparison of slopes.

A graph portrays the simple regression lines before and after the occurrence, and the counterfactual continuation (at or after the occurrence) of the "before" regression line, for comparison with the actual "after" regression line (see graph below).

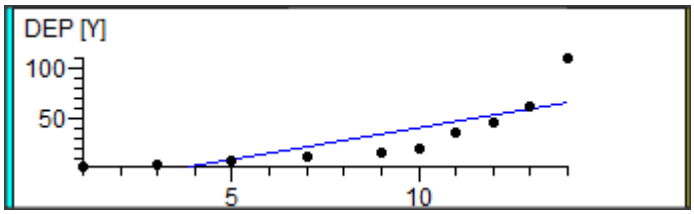
Comparison of regression coefficients

This option permits a comparison of two slope coefficients, e.g those found before and after a public health intervention or other occurrence, controlling for other (and possibly time-related) factors. It can compare regression coefficients for the same variable, based on data relating to different periods, with the same independent variables each time. This option requires two prior regression analyses, so that the two regression coefficients and their standard deviations are known. Welch's *t*-test is used.

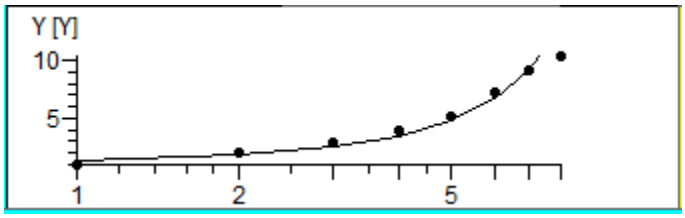
Graphs

For a simple linear regression, five graphs showing the regression lines are displayed. The regression lines are truncated at the edges of the graph. If there are identical values, they are superimposed on each other. The vertical axis refers to the dependent variable, and the horizontal axis to the independent variable.

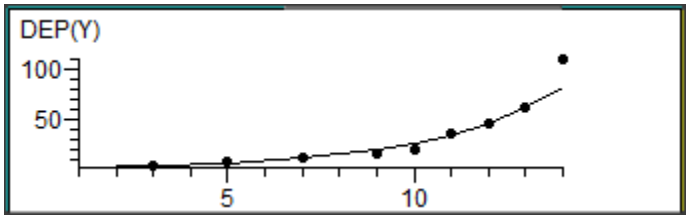
The following graph shows the regression of Y on X:



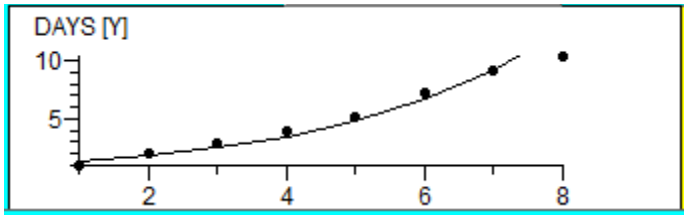
The following graph shows the regression of log Y on X:



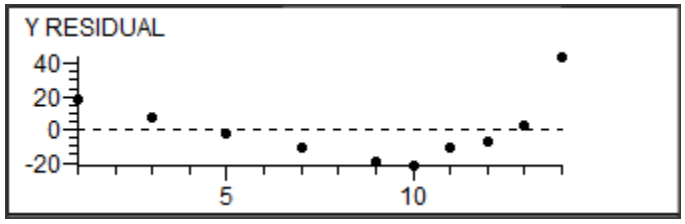
The following graph shows the best-fit exponential curve..



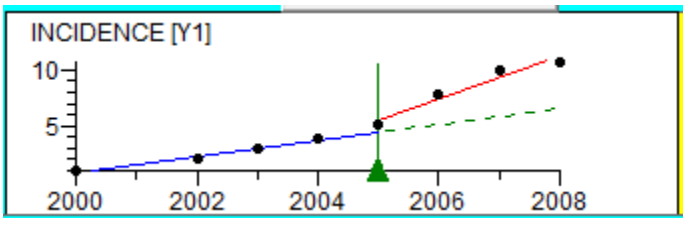
The following graph shows an elasticity curve.



In addition. a scatterplot is displayed, showing the distribution of residuals (the observed value of the independent variable minus the value computed from the simple linear regression equation).



If the interrupted time series regression option is selected, the following graph is displayed. The vertical line represents the time of occurrence, and the blue and red lines refer to the periods before and after this time. The dotted green line represents the counterfactual continuation of the "before" regression line.



METHODS

Note: The program's accuracy has been validated against the certified results for the statistical reference multilinear regression dataset provided by the National Institute of Standards and Technology (http://www.itl.nist.gov/div898/strd/).

Regression equation

The regression equation is computed by the usual formulae (as listed by, for example, Yeomans 1970: 201-205). The log transformations use natural logs. Log transformations are used in the calculation of the exponential Waner 2008) and constant elasticity (Pedace 2013) curves.

Two-tailed P values for the partial regression coefficients are computed from the inverse of *S* (Kymn 1970) which follows an *F* distribution with (*N* - *k*, *N* - *k*) degrees of freedom,

where $S = (1 + r) / (1 - r)$

r = the corresponding partial correlation coefficient

n = size of sample

k = number of variables

$F = 1 / \{[(1 + \text{abs}(r)] / [1 - \text{abs}())]\}$

The standard error of the estimate is the square root of the residual mean square. The *F* value to test the significance of the model is the ratio of the regression mean square to the residual mean square.

Interaction terms

The product of the values of the two variables involved in each interaction is treated as an additional term in the regression model.

The significance of the difference between the R-squared values before and after inclusion of the interaction(s) is appraised by an *F* test, using the formula

$$F = [(R_2^2 - R_1^2)/(k_2 - k_1)] / [(1 - R_2^2)/(n - k_2 - 1)]$$

where R_2^2 = *R*-square for the second model (the model with the interaction[s])

R_1^2 = *R*-square for the first model (the model without interactions)

n = total sample size

k_2 = number of predictors in the second model

k_1 = number of predictors in the first model

Nonparametric regression analysis

The nonparametric regression analysis procedures are described by Daniel (1995: 622-625), Sprent (1993: 195-202) and Sen (1968). The analysis is not done if there are over 146 values. Three alternative ways of estimating beta (the slope coefficient) are used .

If up to 30 numbers are entered, Theil's estimator (Theil 1950) is computed by a method described by Sprent (1993: 195-198). If more than 30 sets of values are entered, Sen's method (Sen 1968) is used ; but if there are more than 146 different sets the program employs the abbreviated Theil method (Sprent 1993: 198-202), which uses a systematic sample of the data. For the Sprent and abbreviated Theil methods, which (unlike Sen's method) assume distinct values of the independent variable, the program treats tied observations as if they were not identical by imputing differences of (alternately) 0.000001 or -0.000001.

The point estimate of *beta* (ß) is the median value of $ß_{ij}$, where

$$ß_{ij} = (y_j - y_i) / (x_j - x_i)$$

for each pair of values of the independent variable x (x_i and x_j) and the corresponding values of the dependent variable y (y_i and y_j). Using Sprent's method, $ß_{ij}$ is calculated for all of the N(N-1)/2 possible pairs of values; zero values of (x_j - x_i) are changed to 0.000001 or -0.000001 (alternately). In Sen's procedure $ß_{ij}$ is calculated only if (x_j - x_i) is not zero. In the abbreviated Theil procedure, each of the first N/2 pairs in the sequence is then linked with the pair situated N/2 positions further along the array; $ß_{ij}$ is computed only for these linked observations; zero values of (x_j - x_i) are changed to 0.000001 or -0.000001.

Alpha is estimated by two alternative formulae. The first is the median of the (y_i -ßx$_i$) terms for the N pairs of observations, and the second (Daniel 1995: 623-624) is the median of the averages of the (y_i - ßxi) terms calculated for each of the pairwise combinations of observations. . Both estimators are shown if they differ. The first estimator is recommended if deviations from the regression model cannot be assumed to be symmetrical; the second estimator of alpha (which is not calculated if the abbreviated Theil procedure is used) is recommended if the symmetry assumption is tenable.

Confidence intervals for beta are obtained from an array of values of b_{ij} in order of increasing magnitude. Sen's method (Sen 1968) uses critical values provided by a large-sample formula based on a variance estimate corrected for ties, and Sprent's method (Sprent 1993: 199-202), based on Theil's, uses critical values based on the critical value for Kendall's *tau* for significance at the nominal 5% level in two-tailed tests, obtained from Siegel and Castellan (1988: 363, Table RII) and Sprent (1993: Table IX). Approximate confidence intervals are estimated in a similar way in the abbreviated Theil procedure, using critical values based on formula 2.3 in Sprent (1993: 34).

Coefficients of variation

R-squared (R^2) is the ratio of the regression sum of squares to the total sum of squares.

The formula for the adjusted coefficient of determination (Zar 1998: formula 20.23) is

$$1 - [(n - 1) / (n - m - 1)] / (1 - R^2)$$

where n = sample size
m = no. of independent variables

Effect of removing a variable

The significance of the change in *R*-squared resulting from the removal of a variable is assessed by *partial F*, with degrees of freedom *df1* and *df2*.

Partial F = (*RSSp - RSSq*) / (*df1 / df2 * RSSq)*

where *RSSq* = residual sum of squares in the larger model
RSSp = residual sum of squares in the smaller model
sf1 = degrees of freedom for *RSSp,* minus degrees of freedom for *RSSq*
df2 = degrees of freedom for *RSSq*

Outliers

Outliers are defined as cases where the standardized residual (the difference between the observed and predicted values of the dependent variable, divided by the standard error of the estimated) is 2 or more.

Correlation coefficients

The zero-order and partial coefficients are computed by the usual formulae (as listed by, for example, Daniel 1995: 391-393 and 446; or Yeomans 1970: 179 and 197-205).

The significance of zero-order coefficients is assessed by a *t* test (Daniel 1995: formula 9.7.3):

$$t = r * \sqrt{[(n - 2) / (1 - r^2)]}$$ with n - 2 degrees of freedom,

where r = correlation coefficient
n = sample size

If n is less than 30, a test based on Hotelling's modified *z* transformation is used (Zar 1984: 392, Sokal and Rohlf 1981: 587)

$$z = H\sqrt{(n - 1)}$$

where H = Hotelling's modified *z* transformation of *r*

$= T - (3T + r) / 4n$
T = *z* transformation of r
$= 0.5. \ln[(1 + r) / (1 - r)]$

The significance of partial correlation coefficients is assessed by a *t* test (Daniel 1995: formula 10.6.10):
$t = r * \sqrt{[(n - k - 1) / (1 - r^2)]}$ with $n - k$ -1 degrees of freedom,
where r = partial correlation coefficient
n = sample size
k = number of predictors

Cohen's *f-squared*

Cohen's *f-squared* = *R*-squared / (1 - *R*-squared)

Level-importance

The level-importance of an independent variable is the product of the variable's regression coefficient and the variable's mean value. It is also expressed as a percentage of the mean value of the dependent variable. Interactions are not taken into account.

Exponential regression

As specified by (for example) Waner S (2008), a simple linear regression is performed based on the independent variable X and the log10-transformed dependent variable Y, and *m* and *b* are the intercept and slope of the regression line.
The exponential model is then Y = *A* x (*R* to the power of X).
where $A = 10^m$
$R = 10^b$

G-computation

The module applies the simple procedure described by Snowden *et al*. (2011), and explained in detail in a web appendix to their paper. It uses the multiple regression coefficients to compute two counterfactual (i.e., predicted) values of the dependent (outcome) variable for each subject, based respectively on the presence or absence (observed or imaginary) of exposure to the dichotomous variable of interest. The total set of counterfactuals is then regressed on the value (observed or imaginary) of the binary valuable to obtain an estimate of the marginal effect of the binary variable. This estimate is the mean of the differences between each subject's counterfactual values.

Sample size

The estimates of sample size are based on methods suggested by Harris (1975), Green (1991), and Maxwell (2000) (see text above).

Interrupted time series regression

Welch's t-test is used for the comparisons of slopes and mean values; the Satterthaite-Welch adjustment is used for the degrees of freedom. [See the formulae in the "Welch's *t*-test" article in Wikipedia (https://en.wikipedia.org/wiki/Welch%27s_t-test). The calculated degrees of freedom are rounded down to the nearest integer; if the calculated degree of freedom is less than 1, it is taken as 1.

The *runs test* for serial correlation is based on the direction of the discrepancies between the observed values and the values computed from the regression equation. It compares the number of runs of uninterrupted sequences in the same direction (positive or negative) wit h the number expected in a random sequence. The runs test is described in numerous texts (e.g. Siegel and Castellan1988: 58-64; Zar 1998: 583-585; Sprent 1993: 82-84). If there are <21 values in the sequence, *P* is reported as <0.05, <0.1, <0.2 or >0.2 (or, for one-tailed tests, <0.025, <0.05, <0.1 or

>0.1), using the table of critical values supplied by Zar (1998: App171-App179). In other instances an approximate *P* is computed by formulae 25.14 to 25.16 in Zar (1998: 584).

The *Durbin-Watson test* (Durbin and Watson, 1951) for serial correlation is based on the magnitude of the discrepancies between the observed values and the values computed from the regression equation. The formula is

$$D = \sum[(e_i - e_{i-1})^2] / \sum e_i 2^2]$$

where e_i = the discrepancy for a specific value (other than the first) in the series)

e_{i-1} = the discrepancy for the previous value in the series.

D is compared with tabulated critical values (the lower bound [*DL*] and the upper bound [*DU*]) for $P = 0.2$ (or, for one-tailed tests, 0.1), using the table of critical values supplied by Zar (1998: App171-App179).

K. CONTROLLING AN UNMEASURED CONFOUNDER

This module performs a sensitivity analysis to see how the strength of an observed association with a biunary ("yes-no") variable might be reduced or enhanced by controlling for a hypothetical unmeasured confounder. The calculation is based on scenarios that make different assumptions concerning the strength of the confounder (expressed as an odds ratio or hazard ratio) and its prevalence in groups exposed and unexposed to some factor or (in a case-control study) in cases and controls.

If the adjustment renders the association negligible or nonsignificant, or reverses its direction, and the scenario is a plausible one, this points to a need to measure and take account of other variables, or to be circumspect when drawing conclusions.

The program requires entry of the odds ratio or (for studies that take account of time-to-event) the hazard ratio that expresses the observed association, and its confidence limits (95% or other). These figures may be derived from a Mantel-Haenszel, logistic regression, Cox regression, or other analysis in which allowance was made for the effects of known (measured) variables. Alternative sets of results are provided, depending on whether the prevalence of the unmeasured confounder is to be considered higher in the exposed (or cases) or in the unexposed (or controls).

The unmeasured confounder is assumed to be binary ("yes-no"). It can be regarded as representing a set of unmeasured confounders and their combined effect ("the dichotomy of high risk versus low risk determined by multiple risk factors" – Lin *et al.* 1998).

The computation is based on a procedure described by Lin et al. (1998), who say that it is applicable to any study design, prospective or retrospective, matched or unmatched.

Different scenarios are used, their respective assumptions being that the hypothetical confounder's effect on the outcome variable is expressed by an odds ratio or hazard ratio of 10, 9, 8, 7, 6, 5, 4, 3, 2, 0.5, 0.45, 0.4, 0.35, 0.3, 0,25, 0.2, 0.15, or 0.1), and that the confounder's prevalence is 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% or 100% in one group, and less (between 0% and 90%) in the other). The bounds (i.e., the most extreme effects of the adjustment) are reached when the prevalences are 100% and 0% respectively.

The table of results is extensive, but attention need be directed only at those scenarios (with respect to the hypothetical confounder's prevalences and the odds or hazard ratio expressing the strength of its effect) that are deemed plausible.

The adjusted result is marked with two asterisks if it is nonsignificant (i.e., if 1.0 falls within the confidence interval), and with three asterisks if the adjustment has reversed the direction of the association). If a scenario that appears to be plausible renders the odds ratio or hazard ratio negligible or nonsignificant, or reverses its direction, this points to a need to include other

variables in the analysis or, failing that, for circumspection when drawing conclusions from the the study findings.

The adjusted estimates of the odds ratio or hazard ratio may be termed “externally adjusted” estimates, since the assumptions about the hypothetical confounder's effect on the outcome variable are not based on the study data (Greenland 1996).

The procedure should be a useful one although it is based on various assumptions that are not necessarily met, e.g. that the effect of the confounder is identical in the exposed and unexposed groups, that the confounder is conditionally independent of the exposure variable or other covariates, that hazard functions for the exposed and nonexposed are proportional over time, and that the observed odds ratio for a binary outcome is derived from a log-linear regression analysis. However, simulation studies by Lin *et al*. (1998) show that when applied to unmeasured binary confounders the procedure yields results that are sufficiently accurate to be useful, even when events are not rare.

METHODS

The program uses formulae 2.8 and 2.9 of Lin *et al.* (1998) to adjust odds ratios, and the corresponding formula 3.8 to adjust hazard ratios. In both instances, the observed odds ratio or hazard ratio and each of its confidence limits is adjusted by dividing it by $(R.P_1 + (1 - P_1)) / (R.P_2 + (1 - P_2))$

where R = the assumed effect (odds ratio or hazard ratio) of the unmeasured confounder
= 2, 3, 4, 5, 6, 7, 8, 9, or 10
(or, if the observed effect is negative, 0.5, 0.45, 0.4, 0.35, 0.3, 0,25, 0.2, 0.15, or 0.1)
P_1 and P_2 = assumed prevalences of the confounder in the two groups
(0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% or 100%), where $P_1 > P_2$

L. BAYESIAN ASSESSMENTS OF AN ASSOCIATION

This module is for use by proponents of Bayesian statistics who regard the usual significance tests (tests of null hypotheses of no association) as possibly misleading, and prefer to interpret an observed association by a method that takes account of the pre-study estimate of its strength.

The observed association and the prior estimate of its strength may be expressed as an odds ratio, hazard ratio, rate ratio, or ratio of risks or proportions, as a difference between rates, risks, proportions, or means, or as a standardized difference ("effect size"). The observed measure must be entered, with (if requested) its 90%, 95%, or 99% confidence interval and (optionally) its P value. The direction of the association must be such that the observed effect is positive (i.e., a ratio more than 1, or a difference more than 0).

If a ratio is entered, the module computes a critical value that, if considered plausible, points to the association's credibility, using the CPI (critical prior interval) procedure.

The module also provides a sensitivity analyses, computing Bayes factors for a wide range of prior estimates of strength, extending (for a ratio) from 1.05 to 20 and (for a difference) from one-tenth to twenty times the observed difference.

Optionally, the module also computes Ioannidis's credibility index for the series of pre-study estimates. This index may be helpful in a study in which a very large number of associations is screened, with little prior expectation of finding that an association is true. The pre-study credibility must be entered.
.

CPI (critical prior interval) procedure

This procedure, proposed by Matthews (2001), is based on reverse-Bayes analysis (Greenland 2006), which starts with the posterior result and asks what sort of prior could have led to this result. It shows whether the observed association, when considered together with prior knowledge, can be taken to have credibility at the 90%, 95%, or 99% confidence level.

Specifically, it provides a critical level for the ratio or difference used as a measure of the association. If a measure of that magnitude (or more or less than that magnitude - depending on the direction of the association) is considered to be plausible, in the light of existing (i.e., prior) lnowledge, the association can be regarded as credible.

Bayes factors (sensitivity analysis)

The Bayes factor measures the weight of evidence for the truth of the association, taking account of the prior expectation, following the principle that the lower the expectation, the stronger is the evidence required to demonstrate the truth of the association. A low P value, say Bayesian statisticians, is not necessarily convincing evidence against a null hypothesis (Katki 2008, Goodman 2005); findings with P values near 0.05 tend not to be confirmed in subsequent studies.

Bayes factors are computed for a wide range of prior estimates of strength, extending (for a ratio) from 1.05 to 20 and (for a difference) from one-tenth to twenty times the observed difference, i.e. covering the whole gamut from scepticism to enthusiasm.

The lower the value of the Bayes factor, the stronger is its support for the association. The following guidelines (Jeffreys 1961) are often used :
< 0.010: decisive support for the association
0.010–0.032: very strong support
0.032–0.10: strong support
0.10–0.32: substantial support
0.32–1.00: not worth more than a bare mention
> 1.00: less credible after than before the study

The Bayes factors are estimated by the method described by Ioannidis (2008a). This assumes normality of the effect, and may be inappropriate in small studies.

Credibility index

This index (Ioannidis 2008b) is a measure of the association's credibility (the probability that it is true). It can be used in "discovery-oriented" studies that examine a large number of associations in the expectation that only a very small proportion of them are true. The pre-study odds is arbitrarily set at a default value of 0.001; but in a study where a very large number of associations is examined (e.g. a genome-wide study of genetic associations) this should be replaced by a value as low as 0.000001.

METHODS

If a P value is not entered, it is computed from the confidence interval. For a difference, the method (based on Altman and Bland 2011b) is first to calculate its standard error by dividing the width of the confidence interval by 2A, where A is 1.645, 1.960, or 2.576 (for a 90%, 95%, or 99% confidence interval, respectively), then to calculate the test statistic *z* by dividing the difference by its standard error, and then to derive a two-tailed P value from *z*, using a FORTRAN routine by Hill (1973). For a ratio, the method is the same, but using the natural logs of the ratio and its confidence limits.

CPI (critical prior interval) procedure

The *procedure* is modelled on the interactive Bayesian Credibility Analysis program provided on the Internet at http://statpages.org/bayecred.html. It uses formula 2 of Matthews (2001), but with the 4 in the denominator replaced by 2 x *zed* (see below).

If the confidence interval straddles 1, or if its lower limit is 1, the association is reported to be "not credible".

Bayes factors

The Bayes factors (*B*) are computed by the method described by Ioannidis (2008a, equations 4 and 6).

$$B = \sqrt{(1 + m)} \exp\{(-^{z2}) / [2(1 + 1/m)]\}$$

where $m = \pi A^2 / 2V$

A = the alternative effect (the prior estimate of the ratio or difference).

V = the variance of the observed effect, computed as:

(for the log of a ratio) $V = \{[\ln(H) - \ln(L)] / 2zed\}^2$

(for a difference) $V = [H - L) / 2zed]^2$

(for a standardized difference) $V = [(\text{observed difference}) / z]^2$

H = upper confidence limit at a given confidence level of 90%, 95%, or 99%

L = lower confidence limit at a given confidence level of 90%, 95%, or 99%

z = the *z* statistic derived from the observed P value (e.g., *z* = 2.576 if *P* = 0.01)

zed = 1.645 if confidence level = 90%, 1.96 if confidence level = 95%, and 2.576 if confidence level = 99%.

The results have been checked against a spreadsheet supplied by Ioannidis (2008C).

The computed Bayes factor is not necessarily lowest when the observed effect coincides with A, because A is the *average* prestudy estimate under the assumption that there is a positive effect - it is the average value of a half-normal distribution (Ioannidis, personal communication).

Credibility index

The credibility index is computed by the formula of Ioannidis (2008b). It is expressed as a percentage.

M. OTHER BAYESIAN ASSESSMENTS OF AN ASSOCIATION

This module provides the *conditional error probability*, the *Bayesian false-discovery probability* (BFDP), or both. These two measures are for use by proponents of Bayesian statistics who regard the usual significance tests (which are tests of "no association" null hypotheses) as possibly misleading, on the grounds that they estimate the probability of false reports of an association when there is no true association, rather than pointing to the probable truth or incorrectness of a report that there is an association. A low P value, say Bayesian statisticians, is not necessarily convincing evidence against a null hypothesis (Katki 2008, Goodman 2005); findings with P values near 0.05 tend not to be confirmed in subsequent studies.

Conditional error probability

The conditional error probability (Sellke *et al*. 2001), which is based on a Bayes factor derived from the observed P value, is the approximate lower bound of the posterior probability of the null hypothesis. A value of 20%, for example, means that the association has about a 20% chance, or more, of being spurious. A low value suggests that the observed association is noteworthy .

The computation requires the prior probability that is, a subjective assessment of the probability of the null hypothesis. This assessed probability may be based on prior research, theoretical plausibility, or scientific consensus. The effect of the subjective assessment can be appraised by repeating the program, using different priors.

Bayesian false-discovery probability (BFDP)

The Bayesian false-discovery probability (BFDP) assesses the noteworthiness of an observed association (Wakefield 2007, 2009). It is the approximate probability of the null, and therefore represents the probability of a false discovery (i.e., a false positive report), given the observed odds ratio. A low BFDP indicates that the observed association is noteworthy. The BFDP is influenced by the prior evaluation of the probability that there is an association, following the principle that the lower the expectation, the stronger is the evidence required to demonstrate the truth of the association. The program permits the choice of a number of alternative estimates of this prior probability of an association, and computes a separate BFDP for each alternative.

The computation requires entry of (a) the observed odds ratio and its confidence interval, and (b) an *a priori* specification of the upper limit for the odds ratio, i.e. the level that it is believed unlikely (with a 2.5% probability) to be exceeded

A threshold level for the BFDP is provided, below which the association may be regarded as noteworthy. This threshold is based on the relative costs of false negative reports (false nondiscovery) and false positive reports (false discovery). This necessitates a subjective

decision concerning the ratio of the cost (undesirability) of a false negative report (calling an association non-noteworthy when in fact the association exists) to the cost of a false positive report (reporting an association as noteworthy when in fact the null is true). BFDP results that fall below the threshold, indicating noteworthiness of the association, are marked with an asterisk.

Wakefield (2007) advocates use of the BFDP instead of the false-positive report probability (FPRP), which, according to Lucke (2009), is unsound and can lead to seriously incorrect inferences.

METHODS

Conditional error probability

This is computed by the formula provided by Lucke (2009: p. 149).

Bayesian false-discovery probability (BFDP)

An asymptotic Bayes factor (ABF) is calculated by formula 6 of Wakefield (2007), and multiplied by the prior odds for each assumed probability that there is an association, providing a series of alternative BFDP values, which can be compared with the threshold value. The variance of the log of the odds ratio is derived from the 95% confidence limits, and the prior variance is computed by Wakefield's formula 8 .

The inverse of the normal distribution function (in formula 8) is computed by an adaptation of icnorm, a Delphi unit written by G. MocCormick (http://home.online.no/~pjacklam/notes/invnorm/impl/mccormick/), using an algorithm by P.J. Acklam {http://home.online.no/~pjacklam/notes/invnorm/#Delphi}

The *threshold value* is $R / (1 + R)$, where *R* is the ratio of the cost of a false non-discovery to the cost of a false discovery.

N. COMPARISON OF NUMERICAL DATA IN THREE OR MORE INDEPENDENT SAMPLES

This module compares numerical data (ratio-scale, interval-scale or ordinal scale) in three or more independent samples. It can compare the distributions of ordered categories to which numbers have been allocated.

If a normal distribution is assumed and three to five samples are to be compared, either full data (individual values, or discrete or grouped values with their frequencies) or summary data (means, standard deviations, and size) may be entered for each sample. If more than five samples are to be compared, only summary data may be entered. If full data are entered, the program provides *means and standard deviations*, a *one-way analysis of variance*, a *test for the homogeneity of variances*, measures of the magnitude of the effect (*omega-squared*, *eta-squared*, and *Cohen's f index*), *confidence intervals for the means and for their differences*, t*ests for the differences between means*, and a *test for trend*,. If full data are entered, a covariate can also be entered; the program then provides (in addition) *a one-way analysis of variance on the covariate*, an *analysis of covariance, adjusted means* (controlling for the covariate), *tests for the differences between the adjusted mean*s, and measures of the magnitude of the effect (controlling for the covariate). If summary data are entered, the only results provided are confidence intervals for the means, and tests for the differences between the means.

If a normal distribution is not assumed, only three to five samples may be compared, and full data are required. The program displays the *medians* of each sample, and compares the samples by performing *Mood's median test*, the *Kruskal-Wallis test*, and the *van der Waerden normal-scores test. Pairwise comparisons*, the *Jonckheere-Terpstra test* for trend, and the *Mack-Wolfe umbrella test* for an inverted-U trend are performed.

Analysis of variance

A one-way analysis of variance (single-factor between-subjects ANOVA) is performed. The analysis assumes that the samples were drawn randomly from three to five independent populations with normal distributions and similar variances. A significant result points to a significant difference between the means of at least two of the groups represented.

Levene test for homogeneity of variances

A significant result points to a significant difference between the within-group variances of at least two of the groups represented.

Measures of magnitude of effect

Three measures of the magnitude of the effect – i.e., the strength of the association between the independent variable (represented by the various samples) and the dependent variable – are computed.

Omega-squared (ω^2) is an estimate of the proportion of variability of the dependent variable that is associated with the independent variable (Sheskin 2007: 916-917). By Cohen's criteria, a value of 0.1379 or more indicates a large effect size, 0.0588 or more (but less than 0.1379) indicates a medium effect size, and 0.0099 or more (but less than 0.0588) indicates a small effect size (Sheskin 2007: 917). Cohen (1988) warns that these criteria should be used only when there is no better basis for evaluation. A zero or negative value indicate absence of an association.

Eta-squared (η^2) is an alternative estimate of the proportion of variability of the dependent variable that is associated with differences between the samples; it is a more biased estimate of the population parameter than *omega*-squared, and the program uses an adjusted *eta*-squared, to reduce this bias (Sheskin 2007: 917-918).

Cohen's f index (Sheskin 2007; 918) is a "standard deviation of standardized means". By Cohen's criteria, a value of 0.4 or more indicates a large effect size, 0.25 or more (but less than 0.4) indicates a medium effect size, and 0.1 or more (but less than 0.25) indicates a small effect size.

If a covariate is entered, the measures of magnitude of effect are computed again, controlling for the covariate (Sheskin 2007: 962).

Confidence intervals for the means

If full data are entered, two sets of 90%, 95%, and 99% confidence intervals are computed for the mean of each group. The first set is based on the estimated variance in the specific group, and the second set (which has narrower intervals) is based on the within-groups variance derived from the analysis of variance, and assumes that the variances are homogeneous. If summary data are entered, the second set (which generally has wider intervals) is based on a pooled variance computed as a weighted average of the total variances in the specific groups.

Confidence intervals for the differences between means

The confidence intervals are based on the pooled variance, on the assumption that the variances are homogeneous.

Pairwise comparisons

If a normal distribution is assumed and full data are entered, three tests for the difference between means are performed for each comparison. The first two are simple comparisons, one assuming that the variances are equal, and one not assuming equal variances. These tests are appropriate if the comparison was a planned one, to test an *a priori* hypothesis. The third test, which uses the procedure described by Games and Howell (1976) for pairwise comparisons of

any number of means, takes account of multiple comparisons, and may be used even if there were no *a priori* hypotheses; computer simulations have demonstrated that this procedure is relatively powerful and accurate (Keselman and Rogan 1978). If a normal distribution is assumed and summary data are entered, the second test is omitted. Each group is compared with the first group entered, on the assumption that the first group is a control group The Dunnett (1964) and Tukey-Kramer (Sheskin 2007: p. 973) procedures are employed to take account of multiple comparisons.

If a normal distribution is not assumed, the *Kruskal-Wallis procedure* is used to test the significance of the difference between the mean ranks of the observations in each pair of samples. Two two-tailed P values are computed for each comparison. The first is appropriate if the comparison was a planned one, to test an *a priori* hypothesis. The second test takes account of multiple comparisons by using the Bonferroni-Dunn procedure, and may be used even if the comparison was not planned.

Trend tests

If a normal distribution is assumed, a test for linear trend is performed for the means of the samples (Sheskin 2007: 928-929), with the samples arranged in the sequence in which they are entered (in accordance with a prior prediction). The program reports the P value, the slope – which expresses the average change in the dependent variable that is associated with a change from one sample to the next, and the proportion of the variability of the dependent variable that can be explained by the linear trend.

If a normal distribution is not assumed, the trend of their medians (with the samples arranged in the sequence in which they are entered, in accordance with a prior prediction) is appraised by the *Jonckheere-Terpstra test* for ordered alternatives (Sheskin 2007: 993-1000). The test assumes that the samples were randomly drawn and are independent, and represent populations with distributions that are similar in shape. A one-tailed P value is reported. This is determined from a table applicable to samples with small numbers, or (for numbers not covered in this table, and also for downward trends) by use of a normal approximation.

Umbrella test

The *Mack-Wolfe umbrella test* for an inverted-U trend (Mack and Wolfe 1981) is performed only if there is evidence that, with the samples arranged in the sequence in which they are entered, the values increase and then decrease. The peak sample is specified. If there are two equal peaks or the peak extends over two samples, the left-hand one is chosen. Significance is reported as P <0.01, < 0.05, < 0.10, or >0.10.

Analysis of variance on the covariate

A one-way analysis of variance on the covariate is performed. A significant result points to a significant difference between the means of the covariate in at least two of the groups represented. P values are shown.

Analysis of covariance

An analysis of covariance is performed, showing the total and mean sum of squares for the covariate, as well as the total and mean between-groups and within-groups sums of squares. Two P values are shown. The P value computed for the covariate tests the null hypothesis that there is no correlation between the covariate and the dependent variable; a low P value indicates a significant linear relationship between the covariate and the dependent variable (Sheskin 2007: 956-957). A low between-groups P value points to significant variation of the dependent variable among the samples, controlling for the covariate .

The procedure is described by Sheskin (2007: 953-957).

Adjusted means and their comparison

Adjusted means of the dependent variable (controlling for the covariate) are computed for each sample.

Two tests are performed for each comparison. The first is a simple comparison, appropriate if the comparison was a planned one, to test an *a priori* hypothesis. The second test, which uses *Tukey's HSD* (honestly-significant-difference) procedure, takes account of multiple comparisons, and may be used even if the comparison was not planned.

Mood's median test

The null hypothesis tested by the median test (Mood 1950) is that all the samples come from populations with the same median. This test has poor power, but is very robust against outliers.

Kruskal-Wallis test

The Kruskal-Wallis one-way analysis of variance by ranks tests the null hypothesis that the samples come from populations with the same median. It is based on the assumptions that the samples were drawn randomly from three to five independent populations with distributions that are similar in shape; but it is less affected by differences between the variances than is the parametric single-factor ANOVA (Sheskin 2007: 982). A significant result points to a significant difference between the medians of at least two of the groups represented.

Van der Waerden test

The Van der Waerden normal-scores test (Sheskin 2007: 1007-1019) tests the null hypothesis that the samples represent populations with the same distribution. A significant result points to a difference between at least two of the groups represented.

The advantage of the Van Der Waerden test is that it provides the high efficiency of the standard (parametric) ANOVA analysis when the population is really normal, and has the robustness of the Kruskal-Wallis test when normality assumptions are not satisfied.

METHODS

If grouped values are entered, each observation is allocated the value midway between the lower and upper borders of the group; this may, of course, affect the accuracy of the results.

One-way analysis of variance

The method (based on full information) is described in detail by (*inter alia*) Sheskin (2007: 869-873) and Altman (1991: 218-219).

If only means and S.D.s are entered, the following formulae are used:

$$SSW \text{ (sum of squares within samples)} = \Sigma [SDi^2 * (ni - 1)]$$

$$MSW \text{(mean square within samples)} = SSW / [\Sigma(ni) - k)]$$

$$SSB \text{ (sum of squares between samples)} = \Sigma [(mi)^2 * ni] - \Sigma(mi * ni)$$

$$MSB \text{ (mean square between samples)} = SSB / (k - 1)$$

$$F = MSB / MSW$$

where mi = mean of sample *i*
SDi = standard deviation of sample *i*
ni := size of sample *i*
k = no of samples

Levene test for homogeneity of variances

The method is described by Sheskin (2007: 908-910). It is based on the absolute deviations of the scores from the group means.

Measures of magnitude of effect

These measures are computed by equations 21.41 (for *omega*-squared), 21.44 (for the adjusted *eta*-squared), and 21.46 (for *Cohen's f index*) of Sheskin (2007). Cohen's *f* index is not computed if *omega*-squared is negative.
If a covariate was entered, adjusted values are used when computing these measures (Sheskin 2007: 962).

Confidence intervals for the means

The first set of confidence intervals uses the formula(Sheskin 2007: equation 2.8)

$$Mean \pm t.SE$$

where t = the critical two-tailed value in the t distribution for n - 1 degrees of freedom
SE = standard error of the mean = $SD/\sqrt{n}$
n = size of the sample
SD = standard deviation

If full data are entered, the second set of confidence intervals for the mean uses the formula (Sheskin 2007: equation 21.48)

$$Mean \pm t.\sqrt{(WGMS / N)}$$

where t = the critical two-tailed value in the t distribution for N - 1 degrees of freedom
$WGMS$ = the within-group mean square shown in the ANOVA table (the residual variance)
N = sum of sample sizes

If summary data are entered, WGMS is replaced by the pooled variance, V_{pooled} in the above formula where $V_{pooled} = \sum (Vi * [Ni - 1]) / \sum (ni - 1)$

Confidence intervals for the differences between means

Confidence intervals for the differences between pairs of means are estimated by the formula (Altman 1991: 210):

$$\text{Mean} \pm t. \sqrt{(WGMS)}. \sqrt{(1 / n_1 + 1 / n_2)}$$

where t = the critical two-tailed value in the t distribution for the within-groups degrees of freedom
WGMS = the within-group mean square shown in the ANOVA table (the residual variance)
n_1 and n_2 = sizes of the two samples that are compared

Pairwise comparisons

If a normal distribution is assumed, the simple *t*-tests (for testing *a priori* hypotheses) use formulae 8.7a and 8.11 of Zar (1998). The calculated degrees of freedom for the latter test (formula 8.12) are rounded down to the nearest integer. For the *Games-Howell procedure* (Games and Howell 1976), the program employs formulae 3 and 5 of Toothaker (1993), and appraises significance by comparing the result with critical values for $P < 0.01$ and $P < 0.05$ in the studentized range (Daniel 1995: 702-704 or Sheskin 2007: Table A13).

If a normal distribution is not assumed, formula 22.5 of Sheskin (2007) is used (based on the Kruskal-Wallis test). The Bonferroni-Dunn adjustment is made by multiplying the P value by $s(s-1)/2$, where s = number of samples.

Dunnett's test (Dunnett 1964) and the Tukey-Kramer test (Sheskin 2007 : p. 973) are used for comparisons with a control group.

Trend tests

Formulae for the trend test (assuming a normal distribution) are provided by Sheskin (2007: 928-929). The number of observations (n) used in the formula for *SSlinear* (or *SScomp* in equation 21.17) is the harmonic mean (equation 1.5) of the numbers in the various samples; if the samples are very different in size, use of this mean compromises the accuracy of the analysis (Sheskin 2007: 970). The coefficients required for the analysis are computed by allocating a number (i = 1, 2, 3, etc.) to each successive sample, and then subtracting the mean value of i from each sample's i (coefficient = $i - i_{mean}$). The estimated slope is the sum of the means weighted by the coefficients, divided by the sum of the squared coefficients (Maxwell and Delaney 2004: 248).

The method of calculating the *Jonckheere-Terpstra statistic* is described by Sheskin (2007: 995-996); the normal approximation is computed by Sheskin's formula 12.7. For small numbers (three samples with eight or fewer observations in each), or four or five equally-sized samples with 2 to 5 observations in each), use is made of a table of critical values (Sheskin's Table A24) for one-tailed P values of < 0.005, < 0.01, < 0.025, and < 0.05. This table is appropriate only if the trend is an upward one (Sheskin 2007: 1006).

Umbrella test

The Mack-Wolfe umbrella test with peak unknown, for equal or unequal sample sizes, is described in detail by Hollander and Wolfe (1999: 226-229). The Mack-Wolfe statistic is compared with tabulated critical values (Hollander and Wolfe 1999: Table A.15).

Analysis of variance on the covariate

The procedure is described by Sheskin (2007: 951-953).

Analysis of covariance

The procedure is described by Sheskin (2007: 953-957).

Adjusted means and their comparison

Formulae for the adjusted means and for comparisons of means are provided by Sheskin (2007: 958 and 958-960 respectively). Tukey's HSD test makes use of the studentized range. The number of observations (*n*) used in the formulae is the harmonic mean (equation 1.5) of the sizes of the various samples; if the samples are very different in size, use of this mean compromises the accuracy of the analysis (Sheskin 2007: 970).

Median test

The test is performed by determining the median of the combined samples, and then categorizing the observations (in each sample) that are (respectively) below or above this overall median. If there are observations that are equal to the median, half of them are placed in the "below-median" group and half in the "above-median" group (Sheskin 2007: 646). A chi-square test (with *s*-1 degrees of freedom) is then performed on the resultant 2 x *s* table (where *s* = the number of samples).

Kruskal-Wallis test

The Kruskal-Wallis statistic is computed by formula 22.1 of Sheskin (2007), corrected for ties (formulae 22.3 and 22.4). The statistic is referred to the chi-square distribution, with *s*-1 degrees of freedom (where *s* = number of samples). If the numbers are very small, the P values are approximate.

Van der Waerden test

The van der Waerden chi-square statistic is computed by formula 23.2 of Sheskin (2007).The number of degrees of freedom is *s*-1 (where *s* = number of samples)

O. FACTORIAL-DESIGN AND CROSSOVER TRIALS

This module can analyse factorial-designs that simultaneously evaluate the effect of two factors on a numerical dependent variable, and crossover trials with a numerical dependent variable .

In the *factorial-design study*, each factor can have two or three levels, e.g Treatment and Control, or Treatments A and B and Control. Random allocation of the subjects to the 4, 6 or 9 groups in the study is assumed. The program performs a *between-subjects factorial analysis of variance*, and displays *mean values, with their confidence intervals*. If a factor has three levels, its mean values at different levels are compared, using *Fisher's LSD test*, the *Scheffé test*, and ,*Tukey's HSD test*. *Analyses of the simple effects* of each factor are also performed, and three measures of the magnitude of the effect on the dependent variable are computed (*standard and partial omega-squared*, and *Cohen's f index)*. The *heterogeneity of variances* is tested by the *Brown-Forsythe test* or *Hartley's Fmax test*.

The standard analysis assumes that the samples in the various groups are equal in size. If they are not (e.g. because of loss of subjects), two analyses are performed: one uses the *unweighted-means procedure (*which is suitable for unequal samples), and the other analysis is based on equal-sized samples, after they have been equalized by *deleting randomly-chosen subjects* from the larger group or groups. These are only approximate solutions to the unequal-size problem; but unless the samples are very small or their sizes are very different (in which instances the whole study is of questionable validity), the major results of these two methods may be reasonably similar.

For a *crossover trial* of the effects of two treatments, X and Y, conducted by randomly allocating the subjects to two groups with a different sequence of treatments (X first or Y first), the program performs a *factorial analysis of variance for a mixed design*, and displays *mean values* and *confidence intervals for the mean difference* between treatments (adjusting for sequence). *Analyses of simple effects* are also performed - a separate analysis, in each sequence of treatments, of the effect of treatment, and a separate analysis, for each treatment, of the effect of the order of treatments. The effects of the treatments in the first period are compared, (with confidence intervals for their difference), for use if a "period effect", e.g. a persistent carry-over effect of the previous treatment, is suspected.

The standard procedure used to analyze a crossover trial is appropriate if the numbers in the two sequence groups are equal. If they are not, the program can equalize them by *removing randomly-selected subjects* from the larger group, thus converting it to a smaller but still random sample.

Between-subjects factorial analysis of variance (factorial-design studies)

This analysis of variance assumes a normal distribution in the underlying population, and similar variances in the subgroups. It evaluates the effect of each factor, and the presence of interaction

between them. A significant result for a factor indicates that at least two of the levels of that factor represent populations with different mean values.

The analysis is supplemented by the display of mean values and their 90%, 95%, and 99% confidence intervals, and by tests – *Fisher's LSD (least-significant-difference) test*, the *Scheffé test*, and T*ukey's HSD ((honestly-significant-difference) test* – that compare the means at different levels. These tests are not performed if the factor has only two levels, since the F value shown for the factor in the analysis of variance table then represents the comparison of its two levels. Fisher's LSD test is appropriate for planned tests of a priori hypotheses.

Factorial analysis of variance for a mixed design (crossover trial)

In a crossover study, this analysis of variance deals with the effects of two factors: A, the sequence of the treatments (a between-subjects factor), and B, the specific treatment (a within-subjects factor comparing treatments X and Y, where X and Y may be different treatments, or a treatment and placebo).

The results for factor A represent the effect of the sequence, which may be due to time-related changes, such as growth, seasonal changes, or habituation to the measurement, as well as to a possible carry-over effect of the previous treatment if the "washout period" between the treatments was insufficient. The "between-subjects" result for factor A represents the effect of the sequence without adjustment for the treatment, and the "within-subjects" interaction result for AB represents the effect of the sequence period with adjustment for the treatment (Diaz-Uriarte 2002).

The result for factor B represents the variation attributable to the treatment, adjusting for the effect of the sequence.

The analysis is supplemented by the display of *mean values* (for each treatment in each sequence) and their differences, and 90%, 95%, and 99% confidence intervals for the *mean difference between treatments* (adjusting for the period effect).

A comparison is performed of the effects of the two treatments when they are applied in the first test period; significance is tested and 90%, 95%, and 99% confidence intervals are computed for the difference between their effects. This comparison may be helpful if the results suggest a carryover effect.

Analyses of simple effects

The analyses of simple effects compare the levels of each factor in turn, at a given level of the other factor.

These analyses may be useful if there is significant interaction between the factors.

Measures of magnitude of effect

Three measures of the magnitude of the effect – i.e., the strength of the association between the independent variable (represented by the various samples) and the dependent variable – are computed.

Omega-squared (ω^2) is an estimate of the proportion of variability of the dependent variable that is associated with the two factors and with their interaction (Sheskin 2007: 1146). Two versions are computed – *standard omega-squared*, which assesses the effect on total variability, and *partial omega-squared*, which is said to be more meaningful because variability not attributable to the factor under consideration is eliminated from the total variability. By Cohen's criteria, a value of 0.1379 or more indicates a large effect size, 0.0588 or more (but less than 0.1379) indicates a medium effect size, and 0.0099 or more (but less than 0.0588) indicates a small effect size (Sheskin 2007:1149). A zero or negative value indicate absence of an association.

Cohen's f index (Sheskin 2007; 1149-1150) is a "standard deviation of standardized means". By Cohen's criteria, a value of 0.4 or more indicates a large effect size, 0.25 or more (but less than 0.4) indicates a medium effect size, and 0.1 or more (but less than 0.25) indicates a small effect size .

Cohen (1988) warns that the above criteria should be used only when there is no better basis for evaluation .

Heterogeneity of variances

The analysis of variance is based on assumed homogeneity of the variances. The program usually uses the *Brown-Forsythe test for heterogeneity of variances*, which does not assume normal distributions. If there are only two values in each group, this test is not feasible, and it is replaced by *Hartley's Fmax test* .

A low P value indicates that the variances in the groups are not similar .

It has been suggested that if there is significant heterogeneity, a level lower than 0.05 should be used when evaluating hypotheses based on the analysis of variance (Sheskin 2007: 1144).

Unweighted-means procedure (for unequal sample sizes)

The unweighted-means procedure (Sheskin 2007: 1153-1154, Keppel 1991: 288-291) for analysing a factorial-design study replaces the different sample sizes of the groups with their harmonic mean. The results are roughly equivalent to those of the standard procedure if the differences in sample size are slight, but they are biased – the *F* values derived from the analysis of variance tend to be raised, leading to the suggestion that P values of 0.025 should be required if a 5% level of significance is desired (Keppel 1991: 288). The inaccuracy is less marked if both factors have two levels (Maxwell 2004: 335).

Because of the bias, a standard analysis is also performed, after equalizing the sample sizes by *deleting randomly-chosen subjects* from the larger group or groups, thus converting them to smaller but still random samples; but this obviously lowers the power of the tests. These two analyses may suffice for most purposes.

If the inequality of sample sizes is a reflection of selection bias (e.g. due to a high mortality in one group), neither analysis may be appropriate.

METHODS

Between-subjects factorial analysis of variance, and comparison of means

The method is described by (*inter alia*) Sheskin (2007: 1122-1128, equations 24.1 - 7.27).

The means at different levels of a three-level factor are compared by *Fisher's LSD test* (Sheskin 2007: 1134-1136) and by the *Scheffé test* and *Tukey's HSD test* (using formulae derived from equations 27.38-27.39 and 27.45 respectively). The HSD test uses critical values for $P < .001$, $P < 0.01$, and $P < 0.05$ from Table B5 of Zar 1998.

Factorial analysis of variance for a mixed design

The method is described by Sheskin (2007: 1167-1172).

Confidence intervals for mean values and for differences

Confidence intervals for mean values are estimated by the method described by Sheskin (2007: 1150), using equation 21.48.

In a crossover study, the confidence interval for the mean difference between treatments is estimated by the method described by Sheskin (2007: 174-176; equations 27.74 - 27.81).

The comparison of treatments when they are applied in the first period uses a *t* test for two independent samples (Sheskin 2007: 429 and 1181); the confidence intervals for the difference are estimated by Sheskin's equation 11.17.

Analyses of simple effects

The method is described by Sheskin (2007: 1141-1143).

Measures of magnitude of effect

These measures are computed by equations 27.51-27.53 (for *omega*-squared), 27.57-27.59 (for the adjusted *omega*-squared), and 21.45 (for Cohen's *f* index) of Sheskin (2007). Cohen's *f* index is not computed if *omega*-squared is negative.

Tests for heterogeneity of variances

The *Brown-Forsythe test* is described by Keppel (1991: 102-104) and Sheskin (2007: 910-912).

Hartley's Fmax test, which is based on the ratio of the largest to the smallest group variance, is described by Sheskin (2007: 1143-1144, and 907-908).

Deletion of randomly-selected subjects

For this purpose the program uses a pseudo-random number generator described by Wichman and Hill (1985). Extensive statistical tests have demonstrated the statistical soundness of this algorithm, which derives each number in turn from three seed numbers (in the range 1 – 30,000), which it modifies for subsequent use. Initial values for the seed numbers are generated by Delphi's inbuilt random-number procedures, namely RANDOMIZE, which derives a preliminary seed from the system clock, and RANDOM, which is used to generate three random numbers from which the required seed numbers are computed. Delphi's RANDOM procedure is augmented by an additional randomizing shuffle, using the algorithm of Bays and Durham, as described by Press et al. (1989: 215-217).

The formula for each selection is

$$\text{trunc}(R.M) + 1$$

where R is a random number in the range $0 < R < 1$

M = the original number of subjects in the group.

The same subject may be selected more than once, but previously-selected subjects are filtered out.

Unweighted-means procedure

The unweighted-means procedure is described by Sheskin (2007: 1153-1154) and Keppel (1991: 288-291, 294, and 543).

P. SAMPLE SIZE FOR REGRESSION ANALYSIS

This module estimates the sample size required for a simple or multiple regression analysis, using rules-of-thumb based on the number of predictors (i.e., independent variables) and the expected strength of the association.

The program can report sample sizes for tests of whether R^2 – the *coefficient of determination* (i.e., the square of the multiple correlation coefficient) differs from zero, and of whether a *partial correlation coefficient* (i.e., the correlation between a single predictor and the dependent variable, holding the other predictors constant) differs from zero.

The number of predictors must be entered, together with the expected value of R^2 or the expected value of the partial correlation coefficient, or both these expected values.

Results are presented not only for the entered values of R^2 or the partial correlation coefficient, but for values that that have been suggested (Cohen 1988) as indicative of small, medium, and large effect sizes, and for a very large effect size..

The program uses simple rules-of-thumb to estimate minimal sample sizes for tests with a power of 80% and a significance level of 0.05. The rules are based on the expected strength of the association as well as the number of predictors, and are closer to sample sizes provided by power analytic techniques than earlier rules-of-thumb based only on the number of predictors, such as the rule (Harris 1975) that the required sample size is 50 more than the number of predictors, or rules (Schmidt 1971) that it is 15 to 25 times the number of predictors.

The choice of a power of 80% is based on the idea (Cohen 1988) that typically across the behavioral sciences, a 4 to 1 ratio reflects the relative seriousness of a Type I error to a Type II error, so that if *alpha* = 0.05, the probability of a Type II error should be set at 0..20.

The following values are used as indicative of effect size, both for R^2 and for squared partial correlation coefficients: 0.02 (small), 0.13 (medium), 0.26 (large), and 0.50 (very large).

A rule suggested by Green (1991) is used for R^2. The results agree moderately well with sample sizes determined by power analytic methods. For moderate effect sizes there are no discrepancies exceeding 5% if there are up to 20 predictors. For small effect sizes the rule is reasonably accurate if there are few predictors, but is overestimates the required sample size if there are over 20 predictors. For a large effect size, the sample size is underestimated, but only slightly if there are few predictors.

Rules suggested by Green (19191) and Maxwell (2000) are used for partial correlation coefficients. Their validity depends on the correlations between the predictors. Their formulations and results are similar.

Maxwell *et al.* (2008) point out that these sample sizes may be appropriate if the purpose of the study is to appraise the significance of findings, but may often underestimate or (sometimes) overestimate the sample size required to provide precise estimates of parameters (i.e., with narrow confidence intervals).

Optionally, the program will inflate sample sizes to compensate for the probability that some members of the selected samples will be lost, e.g. because of failure to locate addresses, refusal to participate, or missing data. This requires entry of the expected non-inclusion rate (%). This inflation does of course NOT compensate for possible selection bias .

METHODS

The method used for $\boldsymbol{R^2}$ (Green 1991: page 504) is:
Minimum sample size = L / f^2
where $L = 6.4 + 1.65m - 0.05m^2$
m = no. of predictors
$f^2 = R^2 / (1 - R^2)$

The formulae used for a **partial correlation coefficient** (p) are:
Minimal sample size = $a + m - 1$ (Green 1991: page 507)
where a = 390, 53, or 24 for p values of .02, .13 and .26 respectively
$a = 8 / [p / (1 - p)]$ for other values of p (page 508)
and
Minimal sample size = $[7.85(1 - p)] / p + m - 1$(Maxwell 2000: formula 9)
where m = number of predictors.

If a non-inclusion rate is entered, the program inflates sample sizes by multiplying the computed sample sizes by
$1 / [1 - N / 100)]$
where N = non-inclusion rate %
before rounding them up .

Q. OBTAINING CONFIDENCE INTERVALS FROM A P VALUE, OR VICE VERSA

When two values are compared, this module derives 90%, 95%, or 99% confidence interval of their difference (e.g. between means or proportions) or their ratio (e.g. an odds ratio, risk ratio, rate ratio, or hazard ratio) from a P value, or vice versa.

Confidence intervals from P

This procedure (Altman and Bland 2011a) may be useful when only a difference between two values, or a ratio, has been reported, together with a two-tailed P value but with no confidence interval.

The result should be regarded as approximate, although probably sufficiently precise. Complete correspondence between confidence intervals and the result of a significance test cannot be expected, since alternative statistical procedures yield different confidence intervals, and alternative significance tests do not yield identical P values. Altman and Bland state: "The main context where [the methods] are not correct in small samples where the outcome is continuous and the analysis has been done by a t test or analysis of variance, or the outcome is dichotomous and an exact method has been used for the confidence interval. However, even here the methods will be approximately correct in larger studies with, say, 60 patients or more." The method is not appropriate for comparisons of paired observations.

P from a confidence interval

This procedure (Altman and Bland 2011b) may be useful when a P value is required but has not been reported, or when the reported P value is imprecise (e.g. $P<0.05$). It can be applied to results from meta-analysis and regression analysis. Altman and Bland remark: "we are advocates of confidence intervals as much more useful than P values, but we like to be helpful".)

The procedure has the limitations mentioned above.

METHODS

Confidence intervals from P

For a *difference*, the method (based on Altman and Bland 2011a) is first to calculate the test statistic (z) for a normal distribution test from the P value, then to calculate the standard error (*SE*) by dividing the difference by z, and then

to calculate the lower confidence limit as the difference minus *A.SE* , and the upper confidence limit as the difference plus *A.SE*, where *A* is 1.645, 1.960, or 2.576 (for a 90%, 95%, or 99% confidence interval, respectively).

For a *ratio*, the method is the same, but using the natural log of the ratio and the exponentials (antilogs) of the confidence limits.

P from a confidence interval

For a *difference*, the method (based on Altman and Bland 2011b) is first to calculate its standard error by dividing the width of the confidence interval by 2*A*, where *A* is 1.645, 1.960, or 2.576 (for a 90%, 95%, or 99% confidence interval, respectively), then to calculate the test statistic *z* by dividing the difference by its standard error, and then to derive a two-tailed P value from *z*, using a FORTRAN routine by Hill (1973).

For a *ratio*, the method is the same, but using the natural logs of the ratio and its confidence limits.

REFERENCES

Abdi H (2007) The Bonferroni and Sidak corrections for multiple comparisons. In: Salkin N (ed.) Encyclopedia of measurement and Statistics. Thousand Oaks (CA): Sage.

Abramson JH (2004) WINPEPI (PEPI-for-Windows) computer programs for epidemiologists. Epidemiologic Perspectives & Innovations, 2004, 1:6 (available on the Internet at www.epi-perspectives.com/content/1/1/6).

Abramson JH (2011) WINPEPI updated: computer programs for epidemiologists, and their teaching potential. Epidemiologic Perspectives & Innovations 2011, 8:1 (available on the Internet at /http://archive.biomedcentral.com/1742-5573/content/8/1/1

Abramson JH, Gahlinger PM (2001) Computer programs for epidemiologists: PEPI version 4. Sagebrush Press: Salt Lake City.

Achen CH, 1982. Interpreting and using regression. Sage Publications, Iowa City, pp 71-73.

Agresti A (1980) Generalized odds ratios for ordinal data. Biometrics 36: 59-67.

Agresti A (1990) Categorical data Analysis. New York: John Wiley & Sons.

Agresti A (1996) An introduction to categorical data analysis. New York: John Wiley & Sons.

Agrestii A, Liu I-M (1999) Modeling a categorical variable allowing arbitrarily many category choices. Biometrics 55: 936-94

Aickin M, Gensler H (1996) Adjusting for multiple testing when reporting research results; the Bonferroni vs Holm methods. American Journal of Public Health 86:726-728

Altman DG (1991) Practical statistics for medical research. London: Chapman & Hall.

Altman DG, Bland JM (2006) Treatment allocation by minimisation. British Medical Journal 330: 843.

Altman DG, Bland JM (2011a) How to obtain the confidence interval from a P value. British Medical Journal 343: 2090.

Altman DG, Bland JM (2011b) How to obtain the P value from a confidence interval. British Medical Journal 343: d2304.

Altman DG, Gardner MJ (2000) Regression and correlation. In: Altman DG, Machin D, Bryant TN, Gardner MJ (eds) Statistics with confidence, 2nd edition. BMJ Books.

Amali S, Rolston DE, Fulton AE, Hanson BR, Phene CJ, Oster JD (1997) Soil water variability under subsurface drip and furrow irrigation. Irrigation Science 17 151-155.

Armitage P, Berry G, Matthews JNS (2002) Statistical methods in medical research, 4th edn. Oxford: Blackwell Science.

Battaglia MP, Hoaglin DC, Frankel MR (2009) Practical considerations in raking survey data. Survey Practice 2:no. 5

Bender R, Lange S (2001) Adjusting for multiple testing – when and how? Journal of Clinical Epidemiology 54: 343.

Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. Behavioural Brain Research 125: 279-284.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society series B 57: 289-300.

Benjamini Y, Liu W (1999) A step-down multiple hypothesis testing procedure that controls the false discovery rate under independence. Journal of Statistical Planning and Inference 82: 163-170.

Bernal JL, Cumminsd S, Gasspaarrini A (2016) Interrupted time series regression for the evaluation of public health intervention: a tutorial. International Journal of Epidemiology (advance access) doi: 10.1093/ije/dyw098.

Blalock HM Jr (1979) Social Statistics, revised 2nd edn. New York: McGraw-Hill.

Bilder CR, Loughlin TM (2004) Testing for marginal independence between two categorical variables with multiple responses. Biometrics 60: 241-248.

Bland JM, Altman DG (1997) Cronbach's alpha. British Medical Journal 314: 572.

Campbell UB, Gatto NM, Schwartz S (2005) Distributional interaction: interpretational problems when using incidence odds ratios to assess interaction. Epidemiologic Perspectives and Innovations 2: 1.

Cochran WG (1954) Some methods for strengthening the common chi- square tests. Biometrics 10: 417-451.

Cochrane, D., Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. Journal of the American Statistical Association 44 (245): 32–61.

Cochran WG (1977) Sampling techniques, 3rd edn. New York: John Wiley & Sons .

Cohen JE (1988) Statistical power analysis for the behavioral sciences, 2nd edn, revised. Lawrence Erlbaum Associates.

Cole P (1979) The evolving case-control study. Journal of Chronic Diseases 32:15-27

Cortina-Borja M, Smith AD, Combarros DJ O, Lehmann DJ (2009) The synergy factor: a statistic to measure interactions in complex diseases. BMC Research Notes, 2:105

Cronbach LJ (2004) My current thoughts on coefficient alpha and successor procedures. Educational and Psychological Measurement 64: 391-418.

Croxton FE, Cowden DJ (1939) Applied General Statistics. New York: Prentice-Hall.

Curran-Everett D (2000) Multiple comparisons: philosophies and illustrations. American Journal of Physiology: Regulatory, Integrative and Comparative Physiology 279: R1-R8.

Daniel WW (1995) Biostatistics: a foundation for analysis in the health sciences, 6th edn. New York: John Wiley & Sons.

De Gonzalez AB, Cox DR (2005) Additive and multiplicative models for the joint effect of two risk factors. Biostatistics 6: 1-9.

Decady YJ, Thomas DR (2000) A simple test of association for contingency tables with multiple column responses. Biometrics 56: 893-896.

Diaz-Uriarte R (2002) Incorrect analysis of cross-over trials in animal behaviour research. Animal Behaviour 63: 815-822.

Digby P G N (1983) Approximating the tetrachoric correlation coefficient. Biometrics 39: 753-757.

Donner A (1989) Statistical methods in ophthalmology: an adjusted chi-square approach. Biometrics 45: 605-611.

Dunn OJ (1964) Multiple comparisons using rank sums. Technometrics 6: 241-252.

Dunn OJ, Clark VA (1969) Correlation coefficients measured on the same individuals. Journal of the American Statistical Association 64: 366-377.

Dunnett CW (1964) New Tables for Multiple Comparisons with a Control. Biometrics 20: 482-491

Edwardes MDdeB, Baltzan M (2000) The generalization of the odds ratio, risk ratio and risk difference to r x k tables. Statistics in Medicine 19: 1901-1914.

Edwards JH, Edwards AWF(1984) Approximating the tetrachoric correlation coefficient Biometrics 40: 563.

Edwardes, M.D.D., & Baltzan, M. (2000). The generalization of the odds ratio, risk ratio and risk Difference to rxk tables. Stat. Med. 19, 1901–1914.

Efird J (2011) Blocked randomization with randomly selected block size. International Journal of Environmental Research and Public Health 8: 15-20.

Emerson JD, Wong GY (1985) Resistant nonadditive fits for two-way tables. In: Exploring data tables, trends and shapes (Hoaglin DC, Mosteller F, Tukey JW, eds.) Wiley, New York, pp. 59-86.

Everitt BS (1977) The analysis of contingency tables. London: Chapman and Hall.

Feise D J (2002) Do multiple outcome measures require p-value adjustment? BMC Medical Research Methodology 2: 8

Fieller EC, Hartley HO, Pearson ES (1957) Tests for rank correlation coefficients. I. Biometrika 44:470-481.

Fieller EC, Hartley HO, Pearson ES (1961) Tests for rank correlation coefficients. II. Biometrika 48:29-40.

Finner H (1990) Some new inequalities for the range distribution with application to the determination of optimum significance levels of multiple range tests. Journal of the American Statistical Association 85:191-194.

Finner H (1993) On a monotonicity problem in step-down multiple test procedures. Journal of the American Statistical Association 88:920-923 .

Fleiss JL, Levin B, Paik MC (2003) Statistical methods for rates and proportions, 3rd edn. John Wiley & Sons.

Ford RN (1954) A rapid scoring procedure for scaling attitude questions. In: Sociological Studies in Scale Analysis (Riley, MW, Riley JW Jr, Toby J, eds), New Brunswick, N.J.: Rutgers University Press.

Games PA, Howell JF (1976) Pairwise multiple comparison procedures with unequal N's and/or variances. Journal of Educational Statistics1: 113-125.

Goodman SN (2005) Introduction to Bayesian methods: I. measuring the strength of evidence. Cliical Trials 2: 282-290.

GraphPad Statistics Guide (2013). Available on the Internet at www.graphpad.com/guides/prism/6/statistics/index.htm?stat_the_method_of_bonferroni.htm

Green SM (1991) How many subjects does it take to do a regression analysis? Multivariate Behavioral Research 26: 499-510.

Greenland S (1996) Basic methods for sensitivity analysis of biases. International Journal of Epidemiology 25: 1107-1116.

Greenland S (2006) Bayesian perspectives for epidemiological research: I. foundations and basic methods. International Journal of Epidemiology 35: 765-775.

.Guilford JP, Fruchter B (1986) Fundamental Statistics in Psychology and Education, 6th edition, McGraw-Hill.

Gunther A, Hofler M (2006) Different results on tetrachorical correlations in Mplus and Stata – Stata announces modified procedure. International Journal of Methods in Psychiatric Research 15: 157-166.

Hankins M (2007) Questionnaire discrimination: (re)-introducing coefficient delta. BMC Medical Research Methodology 7: 19.

Harris RJ (1975) A primer of multivariate statistics. New York: Academic Press
.
Heo M, Kim N, Faith FS (2015) Statistical power as a function of Cronbach alpha of instrument questionnaire items. BMC Medical Research Methodology 2015, 15:86

Hill ID (1973) Algorithm AS 66. The normal integral. Applied Statistics 22: 424-427.

Hittner JB, May, KM, Silver NC (2003) A Monte Carlo evaluation of tests for comparing dependent correlations. Journal of General Psychology 130: 149-168.

Hofmann RJ (1979) On testing a Guttman scale for significance. Educational and Psychological Management 39: 297-301.

Hogan MD, Kupper LL, Most BM, Haseman JK (1978) Alternatives to Rothman's approach for assessing synergism (or antagonism) in cohort studies. American Journal of Epidemiology 108: 60-67.

Hollander M, Wolfe DA (1999) Nonparametric statistical methods. New York: John Wiley & Sons.

Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6:65-70 .

Hommel G (1988) A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75: 383-386 .

Howell DC (1997) Statistical methods for psychology, 4th edn. Belmont, CA: Duxbury Press.

Ioannidis JPA (2008a) Effect of formal statistical significance on the credibility of observational associations. American Journal of Epidemiology, 168: 374-383.

Ioannidis JPA (2008b) Calibration of credibility of agnostic genome-wide associations. American Journal of Medical Genetics. Part B Neuropsychiatric Genetics 147B: 964-972..

Ioannidis JPA (2008c) Simplebayes.xls; available on the Internet at www.dhe.med.uoi.gr/software.htm

Jacobson PE Jr (1976) *I*ntroduction to Statistical Measures for the Social and Behavioral Sciences, Hinsdale, Ill.: Dryden Press, pp 430-434.

Jeffreys H (1961) Theory of probability, 3rd edition. New York: Oxford University Press.

Kalilani L, Atashili J (2006) Measuring additive interaction using odds ratios. Epidemiologic Perspectives and Innovations 3: 5.

Katki HA (2008) Invited commentary: evidence-based evaluation of p values and Bayes factors. American Journal of Epidemiology 168: 384-388.

Keppel G (1991) Design and analysis: a researcher's handbook, 3rd edn. Upper Saddle River NJ: Prentice-Hall.

Keselman HJ, Rogan JC (1978) A comparison of the modified-Tukey and Scheffe methods of multiple comparisons for pairwise contrasts. Journal of the American Statistical Association 73: 47-52.

Kruskal W, Majors R (1989) Concepts of relative importance in recent scientific literature. American Statistician 43: 2-6.

Kymn KO (1970) Statistical test of a partial regression coefficient under zero population correlation. The American Statistician 24: 30-31.

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33: 159-174.

Lee J (1992) A cautionary note on the use of the correlation coefficient. British Journal of Industrial Medicine 49: 526.

Lehmann R (1977) General derivation of partial and multiple rank correlation coefficients. Biometrical Journal 19: 229-236.

Lin D Y, Psaty B M, Kronmal R A (1998) Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 54: 948-963.

Lucke JF (2009) A critique of the false-positive report probability. Genetic Epidemiology 33: 145-150.

Lui K-J, Cumberland WG (2004) Interval estimation of gamma for an R x S table. Psychometrika 69: 275-290.
Maxwell AE (1961) Analysing qualitative data. London: Methuen.

Mack GA, Wolfe HA (1981) *K*-sample rank tests for umbrella alternatives. Journal of the American Statistical Association 76: 175-181.

Matthews RAJ (2001) Methods for assessing the credibility of clinical trial outcomes. Drug Information Journal 35: 1469-1478

Maxwell SE (2000) Sample size and multiple regression analysis. Psychological Methods 5: 434-458.

Maxwell SE, Delaney HD (2004) Designing experiments and analyzing data: a model comparison perspective, 2nd edn. Lawrence Erlbaum Associates.

Maxwell SE, Kelley K, Rausch J R (2008) Sample size planning for statistical power and accuracy in parameter estimation. Annual Reviews of Psychology 59: 537-563.

Meng X-L, Rosenthal R, Rubin DB (1992) Comparing correlated correlation coefficients. Psychological Bulletin 111: 172-175.

Mood AM (1950) Introduction to the theory of statistics. New York: McGraw-Hill.

Nie NH, Hull CH, Jenkins JG. Steinbrenner K, Nent FH (1975) SPSS Statistical Package for the Social Sciences. New York: McGraw-Hill

Olkin I, Finn JD (1995) Correlations redux. Psychological Bulletin 118: 155-164.

Paul SR (1988) Estimation of and testing significance for a common correlation coefficient, Communications in statistics: theory and methods 17: 39-53.

Pedace R (2013) Econometrics for dummies. For Dummies, p. 142.

Perneger T V (1998) What's wrong with Bonferroni adjustments. BMJ 316: 123.

Poole C (1991) Multiple comparisons? No problem! Epidemiology 2: 241-243 .

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1989) Numerical recipes in Pascal: The art of scientific computing. Cambridge: Cambridge University Press .

Raghunathan TE, Rosenthal R, Rubin DB (1996) Comparing correlated but nonoverlapping correlations. Psychological Methods 1: 178-183.

Riley MW (1963) Sociological research: A case approach. Harcourt, Brace & World

Riley, MW, Riley JW Jr, Toby J, eds (1954) Sociological Studies in Scale Analysis, New Brunswick, N.J.: Rutgers University Press.

Rosner B (1982) Statistical methods in ophthalmology: an adjustment for the intraclass correlation between eyes. Biometrics 38: 105-114.

Rothman KJ (1986) Modern epidemiology. Boston: Little, Brown and Company.

Rothman KJ, Greenland S (1998) Modern epidemiology, 2nd edn. Lippincott-Raven .

Satterthwaite, F. E. (1946), An approximate distribution of estimates of variance components.", Biometrics Bulletin 2: 110–114.

Schmidt FL (1971) The relative efficiency of regression in simple unit predictor weights in applied differential psychology. Educational and Psychological Measurement 31: 699-714.

Schuessler KF (1961) A note on statistical significance of scalogram. Sociometry 24: 312-318.

Scott NW, McPherson GC, Ramsay CR, Campbell MK (2002) The method of minimization for allocation to clinical trials: a review. Controlled Clinical Trials 23: 662-674.

Sellke T, Bayarri MJ, Berger JO (2001) Calibration of p values for testing precise null hypotheses. The American Statistician 55: 62-71.

Selvin S (2004) Statistical analysis of epidemiologic data, 3rd edn. New York: Oxford University Press.

Sheskin DJ (2007) Handbook of parametric and nonparametric statistical procedures, 4th edn. Chapman and Hall/CRC.

Siegel S, Castellan NJ Jr (1988) Nonparametric statistics for the 97andomisati sciences, 2nd edn. New York: McGraw-Hill.

Simon R (1979) Restricted randomization designs in clinical trials. Biometrics 35: 503-512.

Skrondal A (2003) Interaction as departure from additivity in case-control studies: a cautionary note. American Journal of Epidemiology158: 251-258.

Snowden JM, Rose S, Mortimer KM (2011) Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. American Journal of Epidemiology 173: 731-738; and web appendix, available at http://aje.oxfordjournals.org

Sokal RF, Rohlf FJ (1981) Biometry, 2nd ed*n*. New York: W.H. Freeman.

Sprent P (1993) Applied nonparametric statistical methods, 2nd edn. London: Chapman & Hall.

Tang N-S, Qiu S-F, Tang M-L, Pei Y-B (2011) Asymptotic confidence interval construction for proportion difference in medical studies with bilateral data. Statistical Methods in Medical Research 20: 233-259.

Tang N-S, Tang M-L, Qiu S-F (2008) Testing the equality of proportions for correlated otolaryngologic data. Computational Statistics and Data Analysis 52: 3719-3729.

Taves DR (1974) Minimization: a new method of assigning patients to treatment and control groups. Clinical Pharmacology and Therapeutics 15: 443-453.

Thomas D, Decady YJ (2004) Testing for association using multiple response survey data: approximate procedures based on the Rao-Scott approach. International Journal of Testing 4: 43-59.

Toothaker LE (1993) Comparison procedures. Sage University Paper Series on Quantitative applications in the social sciences 07-089. Newbury Park: Sage Publications.

Uebersax J S (2000) The tetrachoric and polychoric correlation coefficients. Available on the Internet at http://john-uebersax.com/stat/tetra.htm

Uitenbroek DG. Design, wegen en het designeffect in GGD gezondheidsenquêtes[Design, data weighing and design effects in Dutch regional health surveys] . Tijdschrift voor Gezondheidswetenschappen (TSG). 2009(2): 64-8.

Vansteelandt S, Keiding N (2011) Invited commentary: G-computation - lost in translation? American Journal of Epidemiology 173: 730-742.

Vlach P, Plasil M (undated) Analysis of multiple-response data. Available on the Internet at http://statistika.vse.cz/konference/amse/PDF/Plasil+Vlach.pdf (accessed Jan. 12, 2009).

Volker MA (2006) Reporting effect size estimates in school psychology research. Psychology in the schools 3: 653-671.

Wakefield J (2007) A Bayesian measure of the probability of false discovery in genetic epidemiology studies. American Journal of Human Genetics 81: 208-227.

Wakefield J (2009) Bayes factors for genome-wide association studies: Comparison with P-values. Genetic Epidemiology 33: 79-86

Waner S (2008) Finite mathematics on-line topic: linear and exponential regression. Available on the Internet at http://www.zweigmedia.com/RealWorld/calctopic1/regression.html

Welch, B. L. (1947). "The generalization of "Student's" problem when several different population variances are involved". Biometrika 34: 28-35.

Wichman BA, Hill ID (1985) Algorithm AS183. An efficient and portable pseudo-random number generator. In: Applied Statistics Algorithms (ed. P Griffiths, ID Hill). London: Ellis-Horwood Ltd for the Royal Statistical Society.

Williams K (1976) The failure of Pearson's goodness of fit statistic. The Statistician 25:49.

Wright PS (1992) Adjusted p-values for simultaneous inference. Biometrics 48: 1005-1013.

Yeomans K A (1969) Statistics for the social scientist: 2. Applied statistics . Harmondsworth: Penguin Books.

Zaiontz C (2015) Real statistics using Excel. Available on the Internet at http://www.real-statistics.com/matrices-and-iterative-procedures/iterative-proportional-fitting-procedure-ipfp/

Zar JH (1998) Biostatistical Analysis, 4th edn. Upper Saddle River, New Jersey: Prentice-Hall.

Zou GY (2007) Toward using confidence intervals to compare correlations. Psychological Methods 12: 399-413.

Zou GY (2008) On the estimation of additive interaction by use of the four-by-two table and beyond. American Journal of Epidemiology 168: 212-214

Zou GY (2012) Sample size formulae for estimating intraclass correlation coefficients with precision and assurance. Statistics in Medicine 31: 3972-3981.