Matlab Toolbox 'Measures of Effect Size'

Harald Hentschke
Section of Experimental Anesthesiology
Department of Anesthesiology
University Hospital of Tübingen
72076 Tübingen
Germany
harald.hentschke@uni-tuebingen.de

Maik C. Stüttgen
Institute of Pathophysiology & Focus Program Translational Neurosciences
University Medical Center Mainz
55128 Mainz
Germany
maik.stuettgen@uni-mainz.de

This document is a guide and reference manual for the Matlab toolbox 'Measures of Effect Size' abbreviated in the following as MES. First, we provide an overview of the toolbox – its contents, practical aspects of usage, and some implementation detail. Where deemed helpful, brief introductions to some theoretical statistical concepts are included. In the reference section, all implemented measures of effect size are listed, including formulae and examples to illustrate their use. The examples are based on simulated data which are included in the program package.

Sources of information, recommended literature

Our primary source of information and inspiration was (Kline, 2004). The selection of measures of effect size in the toolbox reflects to a substantial degree the material presented in this work. Other excellent texts on effect size are Grissom and Kim (2012) and Cumming (2012). For the computation of exact confidence intervals we consulted mainly Smithson (2003). For treatments of contrasts beyond the simple analyses implemented here, see Rosenthal and Rosnow (1985) and Rosenthal et al. (2000). Howell (2002) is an excellent source for in-depth explanations of factorial analysis.

The toolbox is accompanied by a paper (Hentschke and Stüttgen, 2011) in which we first outline the rationale for complementing or substituting null hypothesis tests by measures of effect size. We wish to stress that this is a reiteration of appeals for a more rational use of statistics put forward time and again by numerous authors over the past decades (see references therein). In the second part, we provide real-life examples from neurophysiology, and provide an overview of the toolbox.

Version

This manual refers to version 1.6 of the toolbox, released April 2018.

Toolbox content, Matlab version and Matlab toolboxes required

The package contains

- the four main functions computing MES:
  - *mes.m* – for 1-sample and 2-sample data sets (complement: t-test)
  - *mes1way.m* – 1-way data sets (one factor; complement: 1-way ANOVA)
  - *mes2way.m* – 2-way data sets (two factors; complement: 2-way ANOVA)
  - *mestab.m* – data tables of categorical outcomes (complement: e.g. chi square test)
- a function for iterative determination of the noncentrality parameter (ncp) of noncentral X^2 (chi square), t- or F-distributions needed for the construction of confidence intervals:
  - *ncpci.m*
- a function for plotting 2-way data sets:
  - mesdplot.m
- a handful of accessory functions

- simulated data used in the example calculations:
    - *exampleData.mat*

The code works on Matlab version 7.5 (Release 2007b) and above. Compatibility with lesser versions has not been tested. The Matlab Statistics Toolbox is required.

Example data and calculations

MES are illustrated with simulated data from a canonical study, accessible in *exampleData.mat*. In this hypothetical study, subjects (student sample) are randomly assigned to one of three groups ('control', 'experimental' and 'experimental 2') and receive a treatment (placebo, and two different kinds of dopamine-reuptake inhibitor). In each of the three groups there are 10 males and 10 females (total N=60). Subjects' performance on a simple reaction time (RT) task is assessed both before and after drug administration, as well as on Conners' continuous performance test (CPT). The CPT assesses sustained attention and impulsivity. Subject's task is to respond to each presentation of a letter on a screen with a button press. However, if that letter reads 'X', the subject is asked to withhold responding. The CPT yields various dependent measures. Of interest here are reaction times as well as the number of errors of commission (pressing the button when X appeared), believed to index impulsivity.

variable name	**size**	**meaning**
group	60*1	0=control, 1=experimental, 2=experimental 2
sex	60*1	0=female, 1=male
iq	60*1	IQ points on a nonverbal scale of intelligence
impulse	60*1	number of points on a self-assessment questionnaire for the personality dimension of "Impulsivity"; range 0 (self-controlled) to 20 (highly impulsive)
rt_pre	60*1	mean reaction time over 100 trials (assessed pre-treatment)
rt_post	60*1	mean reaction time over 100 trials (assessed post-treatment)
com_pre	60*1	number of errors of commission in the CPT assessed pre-treatment
com_post	60*1	number of errors of commission in the CPT assessed post-treatment

Table 1 Data in *exampleData.mat* used for the exemplary computations in the reference section

The organization of the data is designed to facilitate an understanding of the exemplary calculations; the focus is not on efficiency or syntactical elegance. In a similar vein, we wish to emphasize that the example calculations provided for each effect size measure illustrate the functionality of the code, highlight the specifics of the MES and point to conclusions the results suggest. The ensemble of calculations, particularly the sequence in which they are presented, is not meant to exemplify best-practice approaches to analysis.

Independent versus dependent data

In 1-way analyses including the comparison of only two groups, data may be independent (unpaired) or dependent (paired, repeated measures). In analyses of more than one factor, terminology reflects the fact that the data may be dependent along some factors but independent with respect to others. We differentiate between i) completely between-subjects designs (independence along all factors), ii) mixed within-subjects designs (independence along one or more factors and dependence along the other factor(s)), and iii) completely within-subjects designs (dependence along all factors). For the purpose of covering both 1-way and 2-way analyses in the following discussion, we term case i) independent and cases ii) and iii) dependent.

The functions in the toolbox principally accept data of all designs, but for the various MES the results may or may not differ between designs. For some MES, e.g. Hedges's g, design-dependent formulae have been published and are implemented, and it is the user's responsibility to make an informed decision as to which specific formula is most appropriate for the analysis at hand (we provide pointers to relevant literature). For other MES there is no conceptual difference; the question whether data are dependent or not is simply ignored, whereas some MES, like $\psi/S_{D\psi}$ ('psibysd'), by definition exist only for dependent data.

The toolbox functions *mes.m*, *mes1way.m* and *mes2way.m* behave according to the following philosophy:

- if for a given MES different formulae are implemented for dependent and independent data, they are listed in the reference section
- if the MES at hand is 'indifferent' in this respect it will be computed for any design and its value will be independent of the design
- if a MES either conceptually assumes or computationally requires a specific design, a warning or an error is issued if the design varies, depending on the severity of the violation of assumptions
- irrespective of the MES required, ***bootstrapped confidence intervals always depend on the design*** due to the way samples are drawn from the different groups! Please see the paragraph 'Confidence intervals'.

Balanced versus unbalanced data

mes.m, *mes1way.m* and *mes2way.m* principally accept unbalanced data, that is, data in which the number of samples in different groups are unequal. However, there are restrictions and caveats:

i) Naturally, if the data are dependent they must also be balanced.
ii) Particularly the MES computed by *mes1way.m* and *mes2way.m* assume balanced designs. As is detailed in the introductory notes to the 2-way analyses in the reference section, the implementation is geared to data sets which are by design balanced, but in which random and sparse loss of data may occur. The implementation accommodates imbalances in group size, in part by using the harmonic mean of group size as the overall group size or related techniques, but you should be aware that the wider the digression from this assumption the more caution has to be exerted in the interpretation of the results.

Confidence intervals

Confidence intervals (CIs) can be computed in various ways. For a range of MES, we implemented analytical CIs. In general, analytical CIs are straightforward to compute only for simple, mostly unstandardized MES. For example, the confidence interval of the difference between means of two independent groups is

$$CI_{95} = (m_1 - m_2) \pm s_D \, t_{\alpha/2,df} \qquad \text{(eq.1)}$$

where m_1 and m_2 are the means of the first and second group, respectively, s_D is the standard error of the mean difference (see eq.6), and $t_{\alpha/2,df}$ is the critical value of the t statistic which depends on the choice of alpha (usually 0.05) and the degrees of freedom, that is, sample sizes. In Matlab, values of $t_{\alpha/2,df}$ are conveniently obtained from function tinv, and the whole business of computing confidence intervals is usually accomplished in one code line. This way of computing confidence intervals is based on *central* X^2-, t- or F-distributions, which assume the null hypothesis. For mean differences, this approach also works if the null hypothesis is not assumed to be true because the distribution of mean differences does not differ in the two scenarios. The resulting confidence intervals are among those termed 'exact' here. Most measures of effect size, however, have more complex distributions which do differ depending on whether the null hypothesis is assumed to be true or not. Traditional approaches dealing with this difficulty may involve transformation of the effect size statistic to e.g. a normally distributed variable, computation of confidence intervals thereof, and retransformation (see e.g. the explanatory notes for $r_{equivalent}$). The resulting confidence intervals are termed 'approximate'; they are implemented for some MES.

Thanks to the computational power available nowadays, an alternative approach has recently resurfaced which overcomes the quandary of computing confidence intervals which rely on the null hypothesis assumption (Steiger and Fouladi, 1997). Based on so-called noncentral X^2-, t- or F-distributions, it yields 'exact' analytical confidence intervals. Briefly, the noncentrality parameter (ncp) of a noncentral distribution describes the degree of deviation from the null hypothesis. Its value is zero if the null hypothesis is true, and different from zero otherwise. The task of constructing exact CIs consists of an iterative determination of the noncentrality parameter of noncentral X^2-, t-

or F-distributions. This is achieved by *ncpci.m*, which is part of the present package (for an interface-style Windows program to compute ncp, also freely available, see e.g. the Noncentral Distribution Calculator ndc.exe by J. H. Steiger, to be found at http://statpower.net/Software.html). Given the X^2, t or F value and the degrees of freedom in the analysis at hand, the code generates pairs of X^2, t or F probability density functions with ncp as the only free parameter. ncp of each function of the pair is iteratively adjusted such that the X^2, t or F value cuts off the upper and lower 2.5 % of the area under the curves, yielding the lower and upper bounds of the 95 % confidence interval, respectively, of ncp. From these, the confidence intervals of the MES can be computed (Steiger and Fouladi, 1997; Smithson, 2003).

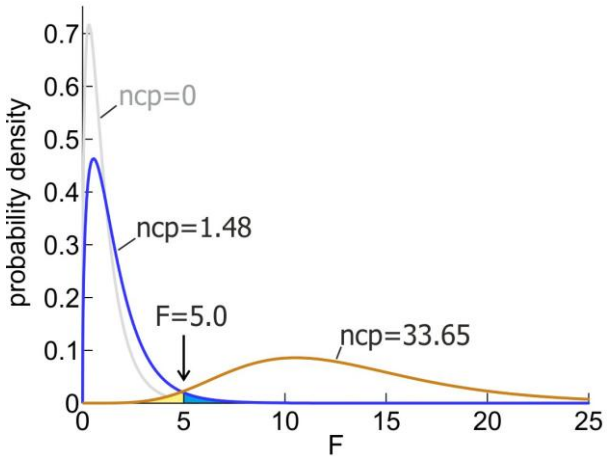

Figure 1. Noncentral F distributions with identical degrees of freedom but varying noncentrality parameters to illustrate the construction of confidence intervals (see main text).

A few points should be noted in connection with confidence intervals and noncentral distributions. First, and foremost, the output of toolbox functions *mes.m*, *mes1way.m* and *mes2way.m* has fields named *xCiType* (replace *x* by the name of the MES computed, e.g. *hedgesg*) which lists the method used for the calculation of CIs. This information has been deemed useful as in an actual computation the method used may be difficult to know because the user's choice may be modified by the code due to the limitations mentioned above. Second, determining ncp is computationally more intensive than much of the code dealing with effect size computation proper; it may in cases rival the computational demand for bootstrapping. Third, if ncp of X^2 or F distributions are needed (as is the case with all 1-way and 2-way MES) and the effect size of the data set in question is small, the lower CIs often cannot be computed via this method. Consider the following example (Figure 1). In a balanced 1-way design with four groups ($df_{numerator}$=3) and ten samples each ($df_{denominator}$=36) let F=5.0. 95% CI of ncp is [1.48 33.65], the lower bound already being close to zero. If F=3.0, the lower bound cannot be computed because this abscissa value is lower than the 97.5th percentile of the 'most central' noncentral F distribution possible, that is, the central F distribution with ncp=0. In such cases, *ncpci.m* will return NaN as the lower CI to indicate this fact; the user may decide to follow common practice by assuming the value to be zero. Bootstrapping may be an alternative (see below); it will yield a lower CI very close to zero. Another solution could be choosing a lower confidence level of e.g. 90%. Finally, it should be mentioned that the paucity of MES for which exact CIs can be computed is partly due to the complexity of the underlying distributions. For example, CIs of η^2 (eta squared) in 2-way analyses are difficult to derive as these statistics follow neither central nor noncentral distributions. Therefore, for all MES except those dealing with tabled data (computed by *mestab.m*), bootstrapping was implemented in the toolbox. Deriving confidence intervals from bootstrapped data is possible for independent and dependent data. When data are independent, samples are drawn from the groups to be compared independently of each other; in the case of dependent data matching pairs (or tuples) are drawn.

Contrasts

In all of the 1-way and 2-way analyses *contrast weights* may be specified. As contrasts may not be widely known, but (may) play a key role in these analyses, the briefest of introductions is provided here; for more information we refer the reader to statistical literature on the subject, e.g. (Kline, 2004) for a succinct introduction and (Rosenthal and Rosnow, 1985;Rosenthal et al., 2000) for authoritative treatments. Consider the example data above, which can be divided into three treatment groups: control, experimental and experimental 2. A contrast, often denoted by Ψ (psi), is a weighted sum of population means. For example, a contrast may be the mean of the control group minus the mean of the experimental group: $\Psi=m_1-m_2$, where m_i is the mean of the i^{th} group. A more sophisticated contrast would be the mean of the first group minus the average of the means of both experimental groups: $\Psi=m_1-0.5*(m_2+m_3)$. More generally, a contrast is the weighted sum of population means:

$$\Psi = \sum_{i=1}^{a} c_i m_i \qquad \text{(eq.2)}$$

where c_i are the individual contrast weights and a is the number of groups (levels of the factor). In the examples just considered, the sets of contrast weights are c=[1 -1 0] and c=[1 -0.5 -0.5], respectively. The contrast weights must sum to zero. Furthermore, the focus in contrast analysis in Kline (2004), which we also pursue here, is on the comparison of means, which means that contrast weights must fulfill additional conditions: if a contrast is to be interpreted as the difference of the means of two subsets of group means, the absolute values of the contrast weights must sum to 2. In this case c is termed a *standard set*. *mes1way.m* will issue a warning if c does not fulfil this requirement. A *standardized contrast* is a contrast divided by a population standard deviation, for example the square root of MS_W, the within-conditions variance pooled across all groups. Thus, like Hedges's g and Glass's Δ, standardized contrasts, here denoted g_ψ, are *standardized mean differences*. They inform us in a readily understandable metric (namely, in terms of the pooled standard deviation of the groups) on the difference between groups. Particularly in conjunction with confidence intervals they permit a comparison of effect magnitude and statistical 'significance' across analyses and studies. In analyses with two or more factors the same holds true, but the computation of contrasts is more varied. Please see the introductory notes to 2-way analyses in the reference section.

Interpretation of effect size magnitude

The guidelines for interpretation of the magnitude of effect size are either based on Cohen (1992) or Kline (2004); for some measures, no guidelines were available, so we decided on our own. Importantly, these guidelines may vary with the type of research conducted, the number of sample points, and a multitude of other factors. See Kline (2004) for a full discussion of this matter.

Overview of optional input arguments into the MES functions

argument	**explanation and possible values**	***Comments***
'isDep'	specifies whether data are dependent (1) or independent (0, the default)	
'missVal'	specifies how to handle missing values (NaNs); if set to 'pairwise', only the missing data point proper is excluded; if set to 'listwise', data points in corresponding positions (i.e. entire row containing any NaN) will be dropped from the analysis. Which of these two settings is applicable and/or the default depends on whether the data are dependent or not. See the individual functions. In *mes2way.m missVal* is not an input argument, data will always be eliminated in pairwise fashion.	*mes.m* and *mes1way.m* only
'nBoot'	determines number of bootstrapping iterations; if given a value n>0, bootstrapping will be performed with n iterations; if 'nBoot' is set to 0, inf or nan, no bootstrapping will be performed, i.e. CIs will be computed analytically, if possible	bootstrapping requires thousands of iterations to yield reliable results
'exactCi'	if 'true', computes exact analytical confidence intervals for effect size measures for which both exact and approximate CIs can be computed; the default is 'false' (exact CIs are based on iterative determination of noncentrality parameters of noncentral t or F distributions, which can be very time-consuming; see documentation for details)	*mes.m* only without effect if bootstrapping is requested
'confLevel'	specifies confidence level for calculation of CI; default is 0.95	
'ROCtBoot'	if 'true', computes bootstrap confidence intervals for the area under the receiver-operator curve according to the 'bootstrap t' method, which is more conservative than the bootstrap percentile method, the default ('false')	*mes.m* only
'trCutoff'	cutoff value for computation of tail ratios, given in units of standard deviations above or below the grand mean; default: 1	*mes.m* only; if set to a positive value, right tail ratio (RTR) will be computed, else left tail ratio (LTR)

'trMeth'	applicable for computation of tail ratios: if set to 'count' (default), ratios will be determined by counting the actual data points beyond the cutoff value; if set to 'analytic', tail ratios will be calculated assuming normal distributions	*mes.m* only
'doPlot'	if 'true' (default) a simple plot for each requested effect size measure will be produced	*mes.m* only
'doDataPlot'	if 'true' a plot of the data reflecting the layout of the analyis will be produced: levels of the factors are in subplot rows/columns, repeated measures data points are plotted with identical colors, backgound colors of the plots reflect contrast weights	*mes2way.m* only
'group'	input argument which must be specified if input data X is a single-column vector (as opposed to a multi-column array in which each column represents one group); g must in this case be a single-column vector of arbitrary numbers coding for the different groups	optional input argument in *mes1way.m*, mandatory input argument in *mes2way.m*
'cWeight'	allows specification of contrast weights for the computation of effect size measures like standardized contrasts and eta squared for focused comparisons; in *mes1way.m* input array c must contain as many columns as there are groups and may contain several rows (=contrasts); in *mes2way.m* only one set of contrast weights may be specified the shape of which depends on the type of contrast to be computed (see specific help therein)	*mes1way.m* and *mes2way.m* only
'tDenom'	for *dependent* data and contrasts, this parameter determines the way F and p values of the contrasts and the confidence intervals of the standardized contrast g_Ψ are computed (please refer to the introductory note to oneway analyses)	*mes1way.m* only

Table 2 Input arguments of the MES functions

Output of the MES functions

The four main functions return a struct array *stats* the fields of which contain the MES, confidence intervals, X^2, t or F statistics as well as additional information like sample size, type of confidence interval, etc. to facilitate post-analysis interpretation of the results. mes1way.m and mes2way.m optionally produce as an additional output the full table of results displayed on the command line. See the detailed help to the individual functions.

Effect size measures covered in *mes.m*

Introductory notes

The 'correlation' type of analyses Users who are not familiar with measures of effect size may wonder why the toolbox offers analyses like point and rank biserial correlations with X and Y unpaired and even of different sample sizes: if the principle of correlation is to evaluate the relations between pairs of data points, how does it work? Let us consider two fictitious variables A and B. Let A be a list of 100 numbers, namely the density of plated neurons after 2 days of cultivation. B would be a list of the same length, but consisting only of ones (some nutritial agent was added) and zeroes (agent was not added). If your data are arranged like this, you can compute point and rank biserial correlations very easily in Matlab. However, chances are that they are not. In most cases, you will have two data columns or lists, one for each of the conditions (with or without agent, respectively), exactly because variable B or, generally speaking, the treatment is a yes-or-no-affair. Thus, you may just plug these two lists into *mes.m*, very much as you would do when running e.g. a t-test for unpaired data. The rearrangement of the data mentioned above is handled within the toolbox, and in case of the point biserial correlation the implementation is different anyways.

The same logic applies to the receiver-operating characteristic analysis: input variable X holds measurements of some parameter in one condition (say, treatment with nutritional agent), Y in the other, and the distinction into false positives and true positives as well as the ROC curve is then computed on the basis of a range of criterion values embracing the minimum and maximum values in the combined data set.

Order of input arguments X and Y For almost all two-sample MES the order of the two samples to be compared matters. In general, X, the first input argument, is assumed to be the first or 'control' group; accordingly, all terms related to this sample are denoted by the subscript '1','X' or 'control'. For example, m_1 or $m_{control}$ denotes the mean of this sample. Accordingly, subscripts '2', 'Y' or 'experimental' (or similar) denote terms of the second group, delivered to *mes.m* via input variable Y. If X and Y are swapped, the results will change. Values for AUROC, the area under the curve of the receiver-operating characteristic, will be symmetrical around 0.5 (e.g. 0.76 for a comparison of X versus Y and 0.24 for the inverse). In the case of Hedges's g, it is solely the sign that changes. For the closely related *Glass's* Δ, the order of X and Y also determines the sign; additionally (and crucially), it affects the magnitude of the result, as only the properties of the first group determine the denominator.

In order to obtain meaningful values (or interpretations) for U_3, CLES, and tail ratios, the user is required to establish - prior to computation - which of the two compared groups is 'higher', that is, which has the higher mean or median. Although this task could be implemented easily, it has not been done: any internal reordering of groups would have to be communicated to the user, and would potentially lead to confusion in the interpretation of results, particularly if reordering depended on the types of effect size measures requested and if multiple comparison (see point below) were made.

Multiple comparisons *mes.m* accepts arrays with more than one column as input variables X and Y. Effect size measures will be computed for each pair of matching columns. The reason for this implementation is that in some analysis situations one may wish to compute effect sizes for a large number of two-sampe data sets, e.g. the bins of matching histogram data. Implementing multiple comparisons in *mes.m* is much more efficient than multiple calls to the function. All other mes functions accept only one data set at a time.

%%

name: ***g_1***

data structure: one sample, metric variable

formula:

$$g_1 = \frac{m - a}{s} \quad \text{(eq.3)}$$

where *m* and *s* are the mean and the standard deviation of the sample, and *a* is a specific comparison value

intuition: standardized distance between the sample mean and a comparison value

complements: one-sample t-test

range: –inf to +inf; no effect: 0

guide for interpretation: small ±0.2, medium ±0.5, large ±0.8

confidence intervals: exact analytical, bootstrap

more information: (Hedges, 1981)

example: IQ is known to correlate with reaction time tasks. Is the average IQ of the sample comparable to that of the general population? IQ tests are standardized to have a population mean of 100 with a standard deviation of 15.

[h,p,ci,stats]=ttest(iq,100) yields t(59)=6.62, $p<10^{-7}$, so the difference turns out to be highly statistically significant. How big is the difference? Application of eq.3 with **mes(iq,100,'g1')** yields g_1=0.85 with a CI95 of [0.56 1.15]. This means that the sample distribution's center (m=110.75) is almost one standard deviation (0.85 sd units) above the population mean, which can be considered a large difference according to the above guidelines. The confidence interval shows that the likely range of the population effect size is from 0.56 to 1.15; so the difference is likely to be no smaller than 0.56 and no larger than 1.15 standard deviations. The sign of g_1 indicates whether the comparison value is above (g_1 positive) or below (g_1 negative) the sample mean.
mes(iq,100,'g1','nBoot',10000) yields bootstrapped CI95 of [0.58 1.22], quite close the analytical values.

In sum, the analysis shows that the sample is not representative of the general population, at least in terms of IQ, which may affect the generalizability of the results.

comments: this measure of effect size is based on *Hedges's g* for two samples

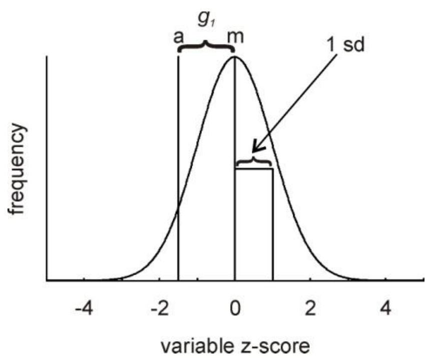

Figure 2. Illustration of *g_1*. Shown is a standard normal distribution, i.e. with a mean (m) of 0 and a standard deviation (sd) of 1. The comparison value, a, is located 1.5 sd to the left of the mean. Accordingly, *g_1*=(0–(–1.5))/1=1.5.

%%
name: ***U3₁***

data structure: one sample, metric or ordinal-scaled variable

formula:

$$U3_1 = \frac{n_{X<a} + 0.5\, n_{X=a}}{n_X} \qquad \text{(eq.4)}$$

where n_X is the number of elements in the sample, and $n_{X<a}$ and $n_{X=a}$ are the number of elements with magnitude smaller than and equal to comparison value *a*, respectively

intuition: fraction of sample below the comparison value

complements: one-sample t-test

range: 0 to 1, no effect: 0.5

guide for interpretation: small 0.4/0.6, medium 0.3/0.7, large 0.2/0.8

confidence intervals: bootstrap

more information: (Cohen, 1988)

example: IQ is known to correlate with reaction time tasks. Is the average IQ of the sample comparable to that of the general population? IQ tests are standardized to have a population mean of 100 with a standard deviation of 15.
[h,p,ci,stats]=ttest(iq,100) yields t(59)=5.35, $p=10^{-8}$, so the difference turns out to be highly statistically significant. How big is the difference? Application of eq.4 with **mes(iq,100,'U3_1','nBoot',10000)** yields $U3_1$=0.22 with a bootstrapped CI95 of [0.12 0.32]. This means that ~22% of the sample values are below the comparison value, rather than 50%, the value expected based on the assumption of a normally distributed variable with a mean of 100. $U3_1$ is symmetric, so in other words, ~78% of the sample values exceed the comparison value. The confidence interval shows that the likely range of the population effect size is from 0.12 to 0.32.
In sum, the analysis shows that the sample is not representative of the general population, at least in terms of IQ, which may affect the generalizability of the results.

comments: this measure of effect size has, to our knowledge not yet been proposed elsewhere and is based on *Cohen's U3* for two samples (see below)

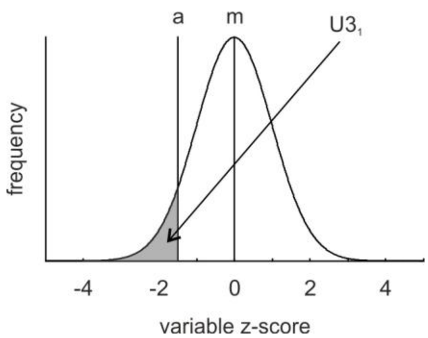

Figure 3. Illustration of $U3_1$. Shown is a standard normal distribution, i.e. with a mean m of 0 and a standard deviation (sd) of 1. The comparison value, a, is located 1.5 sd to the left of the mean. $U3_1$ is indicated by the shaded area, and its value is $U3_1$=0.067.

%%

name: ***mean difference***

data structure: two samples, metric variables

formula:

$$md = m_1 - m_2 \quad \text{(eq.5)}$$

where m_1 and m_2 are the means of the first and second sample, respectively.

For the computation of confidence intervals of *md* the standard error of the mean difference is needed. For two independent samples it is computed as

$$s_D = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{(eq.6)}$$

where s^2_p is the pooled within-groups variance and n_1 and n_2 are the number of cases in groups 1 and 2, respectively. The pooled within-groups variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad \text{(eq.7)}$$

where s_1^2 and s_2^2 are the variances of group 1 and 2, respectively.

For paired samples, computation of standard error of the mean difference takes into account the correlation between the groups:

$$s_D = \sqrt{\frac{s_1^2 + s_2^2 - 2s_1 s_2 r_{12}}{n}} \quad \text{(eq.8)}$$

where s_1 and s_2 are the standard deviations of group 1 and 2, respectively, and r_{12} is the correlation between the groups. However, note that mes.m computes s_D from the standard deviation of the individual difference scores, which is computationally simpler.

intuition: difference between the means of two samples, expressed in the original scale of measurement (i.e. non-standardized)

complements: two-sample t-test

range: –inf to +inf; no effect: 0

guide for interpretation: depends on the scale of measurement

confidence intervals: exact analytical, bootstrap

more information: Kline (2004), Rosenthal et al. (2000)

example: Dopamine activity has frequently been linked to measures of impulsivity. Does administration of APO affect the number of commission errors? **[h,p,ci,stats]=ttest2(com_post(group==0),com_post(group==1))** yields t(38)=–4.01, p=0.0003, i.e. the difference is indeed statistically significant. The mean difference, obtained via **mes(com_post(group==0),com_post(group==1),'md'),** is –9.15 with an exact 95% confidence interval ranging from –13.77 to –4.53. The bootstrapped CI95, obtained via **mes(com_post(group==0),com_post(group==1),'md','nBoot',10000)** is [–14.10 –5.35], quite close to the analytical CI95.

Note that the sign of *md* is dependent on which of the two groups is entered first into mes.m.

comments: this measure of effect size is the most basic one could think of; we included it because in many applications the original units of measurement are meaningful (e.g., millivolts or number of errors) and because in contrast to md itself some users may find its confidence intervals not trivial to compute.

%%

name: ***Hedges's g***

data structure: two independent or dependent samples, metric variable

introductory note: *Hedges's g* is one of several variants of standardized mean differences which are useful and popular measures of effect size for a pair of normally distributed data. However, the nomenclature throughout the literature is very inhomogeneous and confusing: *Hedges's g* also goes by the name *Cohen's d*. This is formally not correct since Cohen used *d* not for a specific implementation of a standardized mean difference, but rather reserved it for the population parameter (see p. 66 in Cohen (1988)), as has been pointed out (Kline, 2004; Grissom and Kim, 2012; Lakens, 2013). Probably to avoid confusion, many authors refrain from associating specific implementations with author names and instead use one symbol for all implementations (typically *d* or δ, occasionally with subscripts), emphasizing that reports of standardized mean differences should always specify the specific underlying formulae. Note that the differences only apply to the standardizer (=denominator s_{within} in eq.9) and the bias correction factor mentioned below; the numerator is always the mean difference. Most of the described standardizers have been implemented in *mes.m*, and the user can determine it via input argument 'esm'. In other words, for practical purposes, we had to associate variants of standardized mean differences with tags (namely, values of input variable esm). For orientation in the case of dependent data, which can be confusing, we provide a list with formulae and values of esm in Table 3, together with detailed references to relevant publications. One may also peruse Grissom and Kim (2012, p. 68 and following) for a detailed account of the inconsistencies in the notation of standardized mean differences.

formula:

$$g = \frac{m_1 - m_2}{s_{within}} \quad \text{(eq.9)}$$

where m_1 is the sample mean of the first group, m_2 is the sample mean of the second group, and s_{within} is the within-groups standard deviation. In the case of **independent** samples, s_{within} is s_p, the pooled standard deviation, that is, the square root of the pooled within-groups variance, weighted by the degrees of freedom in each group:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad \text{(eq.10)}$$

where n_1 and n_2 are the number of cases in groups 1 and 2, respectively, and s_1^2 and s_2^2 are their variances.

Eq.9 is known to yield a biased estimate of the population value. Hedges (1981) described an approximate bias correction formula:

$$g_{unbiased} = g_{biased} c(df_{within}) \quad \text{(eq.11)}$$

where

$$c(df_{within}) = \left[1 - \frac{3}{4df_{within} - 1}\right] \quad \text{(eq.12)}$$

and df_{within} is the degrees of freedom used to compute s_{within}, namely n_1+n_2-2. *mes.m* puts out the bias-corrected *g*.

For **dependent (paired)** data, the approach pursued here is to choose the same denominator as in the case of independent data in order to enable comparisons between studies independent of design. Due to $n_1=n_2=n$ (number of data pairs), s_p reduces to the square root of the averaged variances and is therefore sometimes termed s_{av}:

$$s_{within} = s_{av} = \sqrt{\frac{s_1^{\,2} + s_2^{\,2}}{2}} \qquad \text{(eq.13)}$$

A numerically equivalent way of computing *g* is

$$g = t\sqrt{\frac{2s_D^2}{n(s_1^2 + s_2^2)}} \qquad \text{(eq.14)}$$

where *t* is the t statistic from a t-test for dependent samples, and s^2_D is the variance of the difference scores (for computational reasons, this is in fact the formula implemented in *mes.m*).
For the sake of completeness, it should be mentioned that in another approach, the correlation between the groups is factored into the standardizer (Borenstein, 2009):

$$s_{within} = \frac{s_D}{\sqrt{2(1-r)}} \qquad \text{(eq.15)}$$

where s_D is the standard deviation of the difference scores. However, this approach is not revommended (Bonett, 2015) and therefore not implemented in the toolbox.
There are two more standardizers often encountered in the literature which are implemented in *mes.m* but not listed in this section on Hedges's g as the resulting MES are termed differently in the context of the toolbox, see Glass's Δ and mdbysd below. Again, we point to Table 3 below.
In terms of **bias correction in the dependent case**, information in the literature is scarce and contradictory. As Hedges originally formulated his approximate bias correction only for independent data (Hedges, 1981), it is unclear whether and how this correction also applies to dependent data. While Nakagawa and Cuthill (2007) stick to df_{within} as (n_1+n_2-2) as in the independent case, others set it to $n-1$, where *n* is the number of data pairs (Borenstein, 2009;Cumming, 2012). Cumming (2012) backed his choice with numerical simulations and found that the corrected *g* (*d* in his notation) was closer to the population (true) value than the uncorrected *g*, although some bias remained. The only other explicit reference to a bias correction factor for independent data known to us is in Grissom and Kim (2012), who point to a paper by Bonett (2009). In it, the author states that in a preliminary investigation a correction factor of the form

$$b = \sqrt{\frac{n-2}{n-1}} \qquad \text{(eq.16)}$$

for *g* as computed with s_{av} as the standardizer (cf. eq.13) removed bias. Simulations (https://github.com/hhentschke/simulate-standardized-mean-differences) show that this correction factor is slightly superior to Hedges' c with either definition of df_{within} in many situations. Absent any other work (as far as we are aware) we use Bonett's *b*.

intuition: standardized difference between the sample means, i.e. the difference expressed as the number of standard deviations necessary to move one distribution such that its mean is identical to the other

complements: t-test for two independent or dependent samples

range: –inf to +inf; no effect: 0

guide for interpretation: small: ±0.2, medium: ±0.5, large: ±0.8

confidence intervals: bootstrap (independent & dependent); approximate analytical (independent & dependent), exact analytical (independent)

more information: (Hedges, 1981; Cohen, 1988; Kline, 2004; Nakagawa and Cuthill 2007; Bonett, 2009; Borenstein, 2009; Grissom and Kim, 2012; Cumming, 2012; Lakens, 2013; Bonett, 2015)

example: Dopamine activity has frequently been linked to measures of impulsivity. Does administration of APO affect the number of commission errors?
[h,p,ci,stats]=ttest2(com_post(group==0),com_post(group==1)) yields t(38)=–4.01, p=0.0003, i.e. the difference is indeed statistically significant. *Hedges's g* can be used to quantify the obtained difference, and
mes(com_post(group==0),com_post(group==1),'hedgesg','nBoot',10000) yields *g*= -1.24 with a boostrapped 95% confidence interval ranging from –2.37 to –0.94. The approximate analytical CI95 is [-1.95 -0.59] and can be obtained by leaving out the argument 'nBoot'. The analytical CI can also be computed in an exact fashion (see notes at the beginning of this document and explanations for input argument 'exactCI'). Computation of exact CI may take a longeish time. In the current example, exact CI95 are [-1.94 -0.58], barely different from the approximate CI given above.
Note that the sign of *Hedges's g* is dependent on which of the two groups is entered first into mes.m.

comments: see introductory note

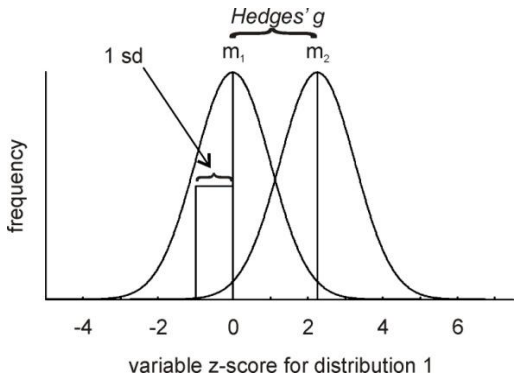

Figure 4. Illustration of *Hedges's g*. Shown are two standard normal distributions, with means m_1 of 0 and m_2 of 2.2, and a standard deviation (sd) of 1 (identical for both distributions). Accordingly, in this example, *Hedges's g*=(2.2–0)/1=2.2.

Formula (without bias correction)	MES Toolbox: input argument 'esm'	(Kline, 2004)	(Nakagawa and Cuthill, 2007)	(Borenstein, 2009)	(Cumming, 2012)	(Grissom and Kim, 2012)	(Lakens, 2013)	(Bonett, 2015)
$g = \frac{m_1 - m_2}{s_p}$ equivalent to $g = \frac{m_1 - m_2}{s_{av}}$ Numerically identical: $g = t\sqrt{\frac{2s_D^2}{n(s_1^2 + s_2^2)}}$	'hedgesg'	Eq. 4.4, 4.5, 4.11	Eq. 1,2 (explicitly only for independ-ent data) Eq. 4 (but note that this is a simplified version of eq.13 here for the case of s_1=s_2 (Kline, 2004))	Eq. 12.11, 12.12 (the latter explicitly only for inde-pendent data)	Eq. 11.9	Eq. 3.5 (explicitly only for independent data)	supposedly eq. 10, but note the error (cf eq. 11.9 in (Cumming, 2012))	Eq. 2
$g = \frac{m_1 - m_2}{s_{rm}}$	*- not implemented because deprecated (Bonett, 2015) -*	n.d.	n.d.	Eq. 12.19-12.20	n.d.	n.d.	Eq. 8, 9	Eq. 1
$g = \frac{m_1 - m_2}{s_1}$	'glassdelta'	4.9	n.d.	n.d.	Eq. 11.8	Eq. 3.1		Eq. 3
$g = \frac{m_1 - m_2}{s_D}$	'mdbysd'	2.15 (4.10)	n.d.	n.d.	Eq. 11.12	(mentioned, p.88)	Eq. 6	Eq. 4

Table 3: Overview of formulae for standardized mean differences for ***dependent (paired, matched)*** data in various publications and their implementation in the MES toolbox. Unless noted otherwise, equation numbers are those used in the cited papers. n.d., not described.

%%
name: ***Glass's*** Δ

data structure: two independent or dependent samples, metric variable

formula:

$$\Delta = \frac{m_{control} - m_{\exp}}{s_{control}} \qquad \text{(eq.17)}$$

where $m_{control}$ is the sample mean of the control group, m_{exp} the sample mean of the experimental group, and $s_{control}$ the standard deviation of the control group. Note that both the sign and the value of *Glass's* Δ depend on which of the two groups is entered first into *mes.m*. The first input argument, X, is assumed to be the control, and the mean difference will be divided by the standard deviation of this group. Like Hedges's g, *Glass's* Δ is multiplied by a correction factor (eq.12) to remove bias for small sample sizes.

intuition: standardized difference between the sample means, i.e. the difference expressed as the number of standard deviations necessary to move one distribution such that its mean is identical to the other

complements: t-test for two independent or dependent samples

range: –inf to +inf; no effect: 0

guide for interpretation: small: ±0.2, medium: ±0.5, large: ±0.8

confidence intervals: bootstrap & approximate analytical

more information: (Hedges, 1981;Glass, 1976)

example: Dopamine activity has frequently been linked to measures of impulsivity. Does administration of APO affect the number of commission errors? **[h,p,ci,stats]=ttest2(com_post(group==0),com_post(group==1))** yields t(38)=–4.01, p=0.0003, i.e. the difference is indeed statistically significant. *Glass's* Δ can be used to quantify the obtained difference, and **mes(com_post(group==0),com_post(group==1),'glassdelta','nBoot',10000)** yields *g*= -5.73 with a boostrapped 95% confidence interval ranging from -10.31 to -3.27. The approximate analytical CI95 is [-7.81 -3.89] and can be obtained by leaving out the argument 'nBoot'.

comments: differs from *Hedges's g* only by the standardizer in the denominator and will frequently yield highly similar results
Glass's Δ has originally been proposed in the context of the evaluation of the effectiveness of different forms of psychotherapy. Since in that context several experimental groups were compared to a single control group, the therapies' effectiveness was best made comparable by using the control group's standard deviation. Under different conditions, *Hedges's g* may be more suitable since it uses both samples' standard deviations to estimate the population's standard deviation. In our example, *Glass's* Δ is much larger than *Hedges's g* because the variance of the control group is considerably smaller than the variance of the experimental group.

%%
name: ***mdbysd***

data structure: two dependent samples, metric variable

formula:

$$mdbysd = \frac{m_1 - m_2}{s_D} \qquad \text{(eq.18)}$$

where m_1 is the sample mean of the first group, m_2 is the sample mean of the second group, and s_D is the standard deviation of the *difference score* (the differences between matching pairs of data in the groups).

intuition: standardized difference between the sample means; will be larger than Hedges's g or Glass's Δ if subjects show high inter-individual variability but consistent effects of e.g. treatment

complements: t-test for dependent samples

range: –inf to +inf; no effect: 0

guide for interpretation: small: ±0.2, medium: ±0.5, large: ±0.8

confidence intervals: bootstrap, exact analytical

more information: (Kline, 2004)

example: As in the previous examples (Hedges's g and Glass's Δ) we ask whether administration of APO affects the number of commission errors, now assuming that the individuals are identical in the control and experimental groups. **mes(com_post(group==0),com_post(group==1),'mdbysd','isDep',1)** yields a value of -0.91 with 95% exact analytical confidence intervals of [-1.43 -0.38]. Thus, there is decidedly a strong effect, the same qualitative result as obtained with Hedges's g and Glass's Δ. The effect of the 'experimental 2' treatment, assessed via **mes(com_post(group==0),com_post(group==2),'mdbysd','isDep',1)** is -1.87 [-2.58 -1.12], more than twice as strong than that of the first experimental treatment.

comments: the standardizer in the denominator is sensitive to correlations between the groups, the advantage being potentially more sensitivity in detecting effects, which must be weighed angainst the disadvantage that the values cannot be compared to Hedges's g or Glass's Δ because the denominator is in a different metric

%%
name: ***$r_{equivalent}$***

data structure: two independent (or dependent) samples, metric variable

formula:

$$r_{equivalent} = \sqrt{\frac{t^2}{t^2 + (N-2)}} \qquad \text{(eq.19)}$$

where *t* is the one-sided t-value of a two-sample t-test, and N is the number of subjects from both conditions
Exact analytical CIs can be computed if the samples are independent. If the samples are dependent or approximate analytical CI95 are preferred, these will be computed. Approximate analytical CIs are computed as for Pearson's correlation coefficient: first, r is transformed to Fisher's Z by Z_r=0.5*log(((1+r)/(1–r))), then the 95% CI of Z_r is found by computing $Z_r \pm t_{df,alpha}/\sqrt{(N-3)}$. $t_{df,alpha}$ is a critical value of the t statistic which depends on the choice of alpha and the degrees of freedom (df). For alpha=0.05 and large n, $t_{df,alpha}$=1.96. 95% CI of $r_{equivalent}$ are obtained by the retransformation r=(exp(2*Z_r)-1)/(exp(2*Z_r)+1)

intuition: correlation between group membership and a metric variable; if the two sample distributions are completely separated, |r|=1, if they are completely indistinguishable, r=0

complements: t-test for two independent or dependent samples

range: –1 to 1; no effect: 0

guide for interpretation: small: 0.2, medium: 0.5, large: 0.8

confidence intervals: bootstrap (independent & dependent); approximate analytical (independent & dependent), exact analytical (independent)

more information: (Rosenthal and Rubin, 2003;Kline, 2004)

example: Dopamine activity has frequently been linked to measures of impulsivity. Does administration of APO affect the number of commission errors?
[h,p,ci,stats]=ttest2(com_post(group==0),com_post(group==1)) yields t(38)=–4.01, p=0.0003, i.e. the difference is indeed statistically significant.
mes(com_post(group==0),com_post(group==1),'requiv','nBoot',10000) yields $r_{equivalent}$=–0.55, with bootstrapped CI95 –0.78 and –0.44. The exact analytical CI95 is [-0.70 -0.28]. It is computed via noncentral F distributions and can be obtained by leaving out the argument 'nBoot' (or setting it to 0) and setting input argument 'exactCi' to 'true' (or 1). If the latter is omitted or set to 'false' (or 0), the approximate analytical CI95 will be computed; it is [-0.74 -0.27], close to the exact values.

comments: also termed point-biserial correlation coefficient; originally introduced as a measure of effect size for meta-analytic research. It is the two-sample case of η (the square root of η^2 (eta squared) explained further below) and can be computed in the two-sample case when only a p- or t-value and the sample size are reported; it draws on the point-biserial correlation, i.e. the Pearson correlation between a metric variable and dichotomous group membership (e.g. control vs. experimental group). It is recommended a) when only sample sizes and p-value are known for a study, b) to complement non-parametric hypothesis tests for which there are no MES, c) when sample sizes are very small or data so nonnormal that computation of other MES such as *Hedges's g* were misleading.

%%

name: ***common-language effect size (CL)***

data structure: two independent or dependent samples, metric variable

formula:

$$CL = \frac{m_1 - m_2}{\sqrt{s_1^2 + s_2^2}} \qquad \text{(eq.20)}$$

where m_1 and m_2 are the means of the two groups, and s_1^2 and s_2^2 are their variances

intuition: CL corresponds to the probability of that a random score from group A will be larger than a random score from group B

complements: t-test for two independent samples

range: 0 to 1; no effect: 0.5

confidence intervals: bootstrap (independent & dependent)

more information: (McGraw and Wong, 1992)

example: Dopamine activity has frequently been linked to measures of impulsivity. Does administration of APO affect the number of commission errors?
[h,p,ci,stats]=ttest2(com_post(group==0),com_post (group==1)) yields t(38)=–4.01, p=0.0003, i.e. the difference is indeed statistically significant.
mes(com_post(group==0),com_post(group==1),'cles','nBoot',10000) yields CL=0.19, with bootstrapped CI95 0.04 to 0.25.

comments: a non-parametric alternative to CL is the area under the ROC-curve (AUROC)

%%

name: ***Cohen's U_1***

data structure: two independent or dependent samples, metric or ordinal-scaled variable

formula:

$$U_1 = \frac{n_{X>\max(Y)} + n_{Y<\min(X)}}{N} \qquad \text{(eq.21)}$$

where $n_{X>max(Y)}$ is the number of elements in group X that are larger than the maximum value of group Y, and $n_{Y<min(X)}$ is the number of elements in group Y that are smaller than the minimum value of group X

intuition: proportion of scores across both groups in areas of non-overlap (see Fig. 5 for illustration); if the distributions are completely separate, U_1=1. If they overlap completely, U_1=0.

complements: t-test for two independent or dependent samples; Mann-Whitney-U-Test; Wilcoxon rank-sum test

range: 0 to 1; no effect: 0, maximum effect: 1

confidence intervals: bootstrap (independent & dependent)

more information: (Cohen, 1988)

example: Dopamine activity has frequently been linked to measures of impulsivity. Does administration of APO affect reaction time?
[h,p,ci,stats]=ttest2(rt_post(group==0),rt_post (group==1)) yields t(38)=–2.03, p=0.0497, i.e. the difference is statistically significant.

Since reaction time distributions are rarely normally distributed, one may opt to choose a nonparametric hypothesis test instead and complete this with a nonparametric measure of effect size, such as U_1.
mes(rt_post(group==0),rt_post(group==1),'U1','nBoot',10000) yields U_1=0.23, with bootstrapped CI95 0.13 to 0.4.

comments: one of the few existing nonparametric MES

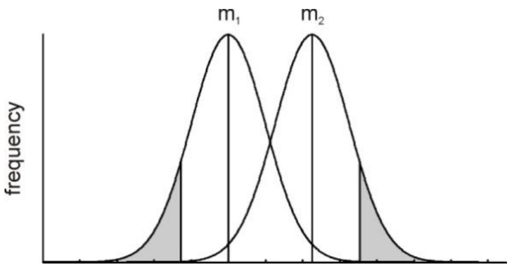

Figure 5. Illustration of *Cohen's U1*. Shown are two standard normal distributions. The gray shaded areas mark the tail regions of the two distributions. The left area is that part of distribution 1 having smaller values than the minimum value of distribution 2. Conversely, the right area is that part of distribution 2 having larger values than the maximum value of distribution 1.

%%
name: ***Cohen's U_3***

data structure: two independent or dependent samples, metric or ordinal-scaled variable

formula:

$$U_3 = \frac{n_{X<median(Y)} + 0.5\, n_{X=median(Y)}}{n_X} \quad \text{(eq.22)}$$

where $n_{X<median(Y)}$ is the number of elements in group X that are exceeded by the median value of group Y, $n_{X<median(Y)}$ is the number of elements in group X that are equal to it, and n_X is the total number of elements in group X

intuition: proportion of scores in group X that are smaller than the typical value (i.e. the median) of group Y; like *U1*, *U3* is best understood as a measure of overlap of two distributions, with overlap being minimal at U3 of 0.5 and maximal at 0 (all elements of group X are above the median of group Y) or 1 (all elements of group X are below the median of group Y)

complements: t-test for two independent or dependent samples; Mann-Whitney-U-Test; Wilcoxon rank-sum test

range: 0 to 1; no effect: 0.5

confidence intervals: bootstrap (independent & dependent)

more information: (Cohen, 1988)

example: Dopamine activity has frequently been linked to measures of impulsivity. Does administration of APO affect reaction time?
[h,p,ci,stats]=ttest2(rt_post(group==0),rt_post(group==1)) yields t(38)=–2.02, p=0.0497, i.e. the difference is statistically significant.
Since reaction time distributions are rarely normally distributed, one may opt to choose a nonparametric hypothesis test instead and complete this with a nonparametric measure of effect size, such as U_3.
mes(rt_post(group==0),rt_post(group==1),'U3','nBoot',10000) yields U_3=0.65, with bootstrapped CI95 0.2 to 1.0. In words, 65 % of the reaction time values in the control (placebo) group are below the median of those in the first treatment group, within (very wide) 95 % confidence margins of 20 to 100 %. Note that the order of input arguments matters in a nontrivial way:

mes(rt_post(group==1),rt_post(group==0),'U3','nBoot',10000) yields U_3=0.45 [0.20 0.65]. When input arguments x and y are swapped, values of U_3 will not be symmetrical around the noneffect value of 0.5 unless the samples in both groups are exactly symmetrically distributed.

comments: one of the few existing nonparametric MES
with *Cohen's* U_3, it is easy to obtain a ceiling or floor effect (note the upper confidence interval in the example above); that is, all values of *X* may be larger than the median of *Y*, but the distributions may still overlap to a considerable degree or not; *Cohen's* U_3 does not differentiate these situations

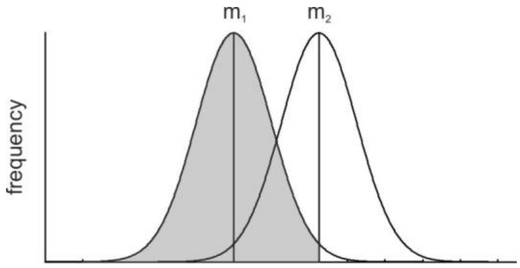

Figure 6. Illustration of *Cohen's U3*. Shown are two standard normal distributions. The gray shaded area marks the proportion of distribution 1 that is exceeded by the median of distribution 2

%%
name: ***AUROC (area under the receiver operating characteristic curve)***

data structure: two independent samples, metric or ordinal-scaled variable

formula: *AUROC* represents the area under the receiver operating characteristic (ROC) curve. This area is found by graphing the data in linear ROC space and then integrating the area below the graph. Formally, successive pairs of hit and false alarm rates (H & F, respectively) are plotted and connected. Then, vertical lines are drawn from each point to the F-axis (abscissa), creating a series of trapezoids. Each of these trapezoids has an area equal to the difference in the F values times the average H value, and the total area is found by summing these areas:

$$AUROC = \frac{1}{2}\sum (F_{i+1} - F_i)(H_{i+1} + H_i) \qquad \text{(eq.23)}$$

where index *i* tracks the pairs of H-F OC points, with (F_1, H_1) being (0,0) and the last point is (1,1).

intuition: like U3, *AUROC* can be understood as a measure of overlap of two distributions, with overlap being minimal at a value of 0.5 and maximal at 0 (all elements of group X are below the minimum of group Y) or 1 (all elements of group X are above the maximum of group Y); *AUROC* also equals the probability that a random score from group X exceeds a random score from group Y

complements: t-test for two independent samples; Mann-Whitney-U-Test; Wilcoxon rank-sum test

range: 0 to 1; no effect: 0.5

confidence intervals: bootstrap, 'bootstrap t' (Obuchowski and Lieber, 1998), analytical (Hanley and McNeil, 1982)

more information: (Bamber, 1975;Stanislaw and Todorov, 1999;MacMillan and Creelman, 2005;McNicol, 2005)

example: Dopamine activity has frequently been linked to measures of impulsivity. Does administration of APO affect reaction time?

[h,p,ci,stats]=ttest2(rt_post(group==0),rt_post(group==1)) yields t(38)=–2.02, p=0.0497, i.e. the difference is statistically significant.
Since reaction time distributions are rarely normally distributed, one may opt to choose a nonparametric hypothesis test instead and complete this with a nonparametric measure of effect size, such as *AUROC*.
mes(rt_post(group==0),rt_post(group==1),'auroc') yields *AUROC*=0.39, with analytical CI95 of [0.22 0.57]. With the classical bootstrap approach via **mes(rt_post(group==0),rt_post(group==1), 'auroc','nBoot',10000)**CI95 range from 0.22 to 0.58. For a more conservative estimate the 'bootstrap t' approach may be used.
mes(rt_post(group==0),rt_post(group==1), 'auroc','nBoot',10000,'ROCtBoot',1) yields CI95 of [0.19 0.58].
Please note that there is no single best way of constructing confidence intervals for AUROC (Obuchowski and Lieber, 1998). When sample sizes are small and AUROC is close to zero or one, confidence intervals in general are very imprecise, the most obvious manifestation being analytical and 'bootstrap t' CI95 ranges exceeding one or zero. The reason is that these estimates assume normality of the pivot statistic, which is often not the case. It may be helpful in such cases to restrict the CI to e.g. 90% or less.

comments: one of the few existing nonparametric measures of effect size
widely used e.g. in psychophysics, machine learning, and engineering
function perfcurve, introduced in the Statistics Toolbox with Matlab Release 2009a, also performs ROC analysis
see also (Jordan et al., 2010), who published a standalone program performing ROC analysis

%%

name: ***RTR, LTR (right tail ratio, left tail ratio)***

data structure: two independent or dependent samples, metric or ordinal-scaled variable

formula:

$$RTR = \frac{p_2}{p_1} = \frac{\dfrac{n_{X>M_{XY}+SD_{XY}}}{n_X}}{\dfrac{n_{Y>M_{XY}+SD_{XY}}}{n_Y}} \qquad \text{(eq.24)}$$

where $n_{X>M_{XY}+SD_{XY}}$ is the number of elements in X that are larger than the grand mean plus the grand standard deviation, $n_{Y>M_{XY}+SD_{XY}}$ is the number of elements in Y that are larger than the grand mean plus the grand standard deviation ('grand' is taken to imply that the measures were computed across the entire sample), n_X and n_Y are the number of elements in groups X and Y, respectively; usually, the larger proportion of scores is taken as the numerator

$$LTR = \frac{\dfrac{n_{X<M_{XY}-SD_{XY}}}{n_X}}{\dfrac{n_{Y<M_{XY}-SD_{XY}}}{n_Y}} \qquad \text{(eq.25)}$$

conventions as in eq.24

intuition: *RTR* is the relative proportion of scores from two different groups that fall in the upper extreme of the *combined* frequency distribution; here, "extreme" means more than one standard deviation above the mean (~84% for a normal distribution), or p_2/p_1; *LTR* is the relative proportion of scores that falls in the lower extreme of the combined frequency distribution (~16% for a normal distribution)

complements: t-test for two independent or dependent samples

range: 1 to +inf; no effect: 1

confidence intervals: bootstrap (independent & dependent)

more information: (Kline, 2004)

example: Dopamine activity has frequently been linked to measures of impulsivity. Does administration of APO affect reaction time?
[h,p,ci,stats]=ttest2(rt_post(group==0),rt_post(group==1)) yields t(38)=–2.02, p=0.0497, i.e. the difference is statistically significant.
Since reaction time distributions are rarely normally distributed, one may opt to choose a nonparametric hypothesis test instead and complete this with a nonparametric measure of effect size, such as *RTR*/*LTR*.
mes(rt_post(group==0),rt_post(group==1),'tailratio','nBoot',10000) yields a tail ratio of zero, with bootstrapped CI95 [0 0.67]. All values are below one, the lower range of possible values defined above. The reason is that the order of input arguments to mes.m matters and the group with lesser values was specified as the first input argument. This results in p2<p1 in equation 18, which is at variance with the standard definition of tail ratios, which assumes the inverse. As stated in the introductory paragraphs, the philosophy of mes is to not swap the order of input arguments internally because this would have to be communicated to the user, unnecessarily complicating interpretation of the data. Instead, the situation can be rectified by exchanging the order of groups in the call to mes.m. However, in the example above p1=0, which will result in infinity. A look at the data with e.g. **hist([rt_post(group==0) rt_post(group==1)])** reveals that reaction times in the treatment group are non-normal with a very large variability compared to the control group. In such a situation the cutoff value may be lowered from the default (one grand standard deviation) to arrive at meaningful values:
mes(rt_post(group==1),rt_post(group==0),'tailratio','nBoot',10000,'trCutoff',0.5) results in a right tail ratio of 7 with CI95 [0.8 9]. In words, the number of samples in the treatment group above [grand mean + 0.5* grand standard deviation] is seven times larger than the number of samples in the control group above this value.
There is an alternative 'analytical' way to calculate tail ratios which assumes an approximately normal distribution of the data in both groups. In the canonical data set, 'impulsivity' qualifies.
mes(impulse(group==0),impulse(group==1),'tailratio',
'nBoot',10000,'trMeth','analytic')
yields a right tail ratio of 1.06, with bootstrapped CI95 [0.19 4.59], demonstrating that both groups are well matched in terms of this characteristic, albeit with a large margin of error.
The left tail ratio can be obtained by specifying a negative factor for the cutoff value:
mes(impulse(group==0),impulse(group==1),'tailratio',
'nBoot',10000,'trMeth','analytic','trCutoff',-1) results in 1.01 [0.84 1.23], confirming the previous result for the lower end of impulsivity values.

comments: widely used e.g. in psychophysics, machine learning, and engineering
The example above illustrates one of the main problems of all nonparametric MES, ceiling and floor effects, and underlines the need to scrutinize data for e.g. normality and/or approximate homogeneity of variability for some MES

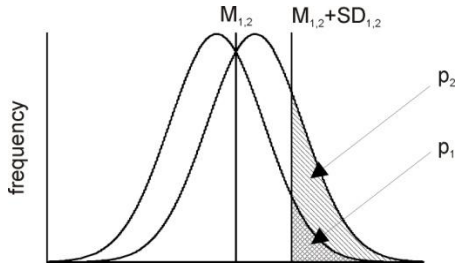

Figure 7. Illustration of *RTR*. Shown are two normal distributions. RTR is computed as p_2/p_1, the area of the left distribution above the cutting point (grand mean $M_{1,2}$ plus one grand standard deviation $SD_{1,2}$) divided by the area of right distribution above that cutting point.

%%%

name: ***r_{rb}*** **rank-biserial correlation**

data structure: two dependent or independent samples, ordinal-scaled variable

formula:

$$r_{rb} = \frac{2}{n_0}\left(\bar{Y}_1 - \frac{n+1}{2}\right) \quad \text{(eq.26) or}$$

$$r_{rb} = \frac{2}{n_1}\left(\frac{n+1}{2} - \bar{Y}_0\right) \quad \text{(eq.27)}$$

where n_0 and n_1 are the number of elements in groups X and Y, respectively, n is the total number of elements, and $\bar{Y}_0$ and $\bar{Y}_1$ are the mean number of ranks in groups X and Y, respectively. The ordinal-scaled variable is transformed to ranks first using tiedranks.m.

intuition: correlation between a ranking (ordinal-scaled variable) and a dichotomy (e.g. experimental group)
how well does group membership predict ranking order?
unlike Pearson's product-moment correlation, r_{rb} measures the degree of monotonicity of a relationship, rather than the degree of a *linear* relationship

complements: t-test for two independent or dependent samples; Mann-Whitney-U test; Wilcoxon rank-sum test

range: –1 to +1; no effect: 0

confidence intervals: bootstrap (independent & dependent)

more information: (Cureton, 1956;Cureton, 1968;Glass, 1966)

example: Continuous performance tests (CPT) are used in the assessment of attention-deficit hyperactivity disorder (ADHD), a symptom of which is impulsive behavior. So, impulsivity is a potential confound when comparing CPT data between groups not matched for impulsivity. Does impulsivity in healthy subjects (as assessed by a personality questionnaire) differ between subjects in the placebo and the experimental groups?
[h,p,ci,stats]=ttest2(impulse(group==0),impulse(group==1)) yields a nonsignificant t(38) of 0.13, with p=0.90. However, absence of evidence (non-significant p-value) does not imply evidence of absence. Quantification of the difference between the two groups can be done by using the rank-biserial correlation coefficient. This coefficient is suitable because impulsivity is measured on an ordinal rather than metric scale level (see (Stevens, 1946)).
mes(impulse(group==0),impulse(group==1),'rbcorr','nBoot',10000) yields a correlation of 0, with CI95 ranging from –0.32 to 0.31. This is reassuring evidence that the groups do not differ appreciably in impulsivity.

comments: one of the few existing nonparametric MES
if the non-dichotomous variable is metric rather than ordinal, the point-biserial correlation coefficient is to be preferred, see above ($r_{equivalent}$).

Effect size measures covered in mes1way.m

Introductory notes

A defining feature of mes1way is the possibility to compute contrast-related MES. Most prominently, this is the standardized contrast g_ψ (g_psi, explained in detail below), which is a oneway equivalent of Hedges's g (explained above). As in the case of Hedges's g, confidence intervals of g_ψ bear a relationship to the p values resulting from t-tests: 95% confidence intervals not enclosing zero correspond to $p<0.05$.

While the computation of the confidence intervals is straightforward for independent data, it merits some deliberation for dependent data. The following note explains the computation of confidence intervals of g_ψ and the t, F and p values of the associated contrast for ***dependent data*** only. It is placed here because it addresses an important general issue.

A contrast is a single degree of freedom comparison between means. Consequently, we may compute t values instead of F values, and the relation $t^2=F$ holds. (Note that mes1way displays only F values in the results table).

The t statistic for a contrast for dependent data, with n subjects and assuming a null hypothesis of zero, is (Kline, 2004)

$$t_{\hat{\Psi}}(n-1) = \frac{\hat{\Psi}}{s_{\hat{\Psi}}} \qquad \text{(eq.28)}$$

where $\hat{\Psi}$ is the contrast and $s_{\hat{\Psi}}$ is the standard error of the contrast. $s_{\hat{\Psi}}$ and the t value are used to compute the approximate confidence intervals of the contrast and the p value associated with the null hypothesis for the contrast, respectively.

The matter of deliberation is $s_{\hat{\Psi}}$, which can be computed in different ways. In one approach,

$$s_{\hat{\Psi}} = \sqrt{MS_{betweenxsubject}\frac{\sum_{i=1}^{a} c_i^2}{n}} \qquad \text{(eq.29)}$$

where $MS_{between\ x\ subject}$ is the between-groups x subjects mean square and the c_i are the contrast weights. Using $MS_{between\ x\ subject}$ as the effect error term is the standard in repeated measures ANOVA and is fine if the data investigated fulfil a number of requirements, among them sphericity. Unfortunately, often they do not. An alternative estimator of $s_{\hat{\Psi}}$ is

$$s_{\hat{\Psi}} = \frac{s_{D_{\hat{\Psi}}}}{\sqrt{n}} \qquad \text{(eq.30)}$$

where $s_{D_{\hat{\Psi}}}$ is the standard deviation of the contrast's difference score. This way of computing $s_{\hat{\Psi}}$ and contrast-associated t statistics has the advantage of not requiring sphericity of the data (Kline, 2004). However, it should be noted that $s_{D_{\hat{\Psi}}}$ takes into account only the variability between the groups compared in the contrast; specifically, groups with a contrast weight of zero do not play a role. It may be argued that this is not exactly in the spirit of best comparability of MES computed from different data layouts, but it must be weighed against the advantage of less stringent requirements on the data mentioned above.

mes1way permits computation of $s_{\hat{\Psi}}$ in both ways. Consequently, F and p values of contrasts and confidence intervals of g_ψ will differ depending on the method chosen. The default is the second method, based on $s_{D_{\hat{\Psi}}}$. As users familiar with ANOVA may expect the other approach, a message alerting the user to this fact is issued on the command line below the table of results. To switch to the method based on $MS_{between\ x\ subject}$, the user must specify optional input variable 'tDenom' as 'msw'. Confidence intervals of the other standardized contrast available, $\psi/S_{D\psi}$, are always based on $s_{D_{\hat{\Psi}}}$.

%%
name: **psi (ψ)**

data structure: two or more independent or dependent samples, metric variable

formula:

$$\Psi = \sum_{i=1}^{a} c_i m_i \qquad \text{(eq.31)}$$

where c_i is the i^{th} contrast weight, m_i the i^{th} group mean, and a the number of groups

intuition: oneway equivalent of mean differences, for example the difference between the mean of a single group and the combined mean from several other groups, expressed in the original units of measurement

complements: analysis of variance & post-hoc tests

range: -inf to +inf; no effect: 0

confidence intervals: exact analytical (independent & dependent), bootstrap (independent & dependent)

more information: Kline, 2004;

examples: 1. Do the control and first treatment group differ in performance (post-treatment error of commission)?
mes1way(com_post,'psi','group',group,'cWeight',[1 -1 0]) yields ψ=-9.15 with exact analytical confidence intervals of [-13.05 -5.25]. In words, subjects in the treatment group committed on average 9.15 errors more than subjects in the control group, our confidence in the assessment (as per the CI95) ranging from 5.25 to 13.05. Reversing the sign of the contrast weights results in identical numerical results with reversed sign. Note that the contrast weights define simply the mean difference between the first two groups; hence, the values of ψ here and of md in the corresponding example calculation above are identical. The confidence intervals are similar, but not identical, because in the oneway calculations here data in the third group come into play for the calculation of the contrast's standard error, on which the confidence interval is based.
2. Does the control group differ from the ensemble of both treatment groups (equally weighted) in terms of performance?
mes1way(com_post,'psi','group',group,'cWeight',[1 -0.5 -0.5]) yields ψ=-6.53 (CI95 [-9.90 -3.15]). The fact that now the value of ψ is lower than in the first comparison implies that the effect in the second treatment group is somewhat weaker than in the first treatment group.
3. The same question as in example 2, but now assuming dependent samples: ψ=-6.53 as before, because the definition of ψ is identical in the dependent and independent case. However, the confidence intervals, [-8.98 -4.07], are slightly more narrow than under the assumption of independence, which is to be expected as correlations between subjects were built into the data set.
Please note that the output of *mes1way.m* (**stats.psi**) will be a column array in which the first element is always NaN: the first position in all output arrays is reserved for the omnibus effect, which contrasts by definition are not. From the second row on, **stats.psi** will hold the values for the contrasts (as many rows as contrasts are specified). The same applies to **stats.psiCi**, the confidence intervals.

comments: see the introductory paragraph on contrasts in this document. Also, once again it shall be noted that in some applications the real units attached to unstandardized statistics may convey a better idea of the difference between groups, the drawback being that the numerical values cannot readily be compared to unstandardized differences derived in other contexts.

%%

name: **g_psi (g_ψ)**

data structure: two or more independent or dependent samples, metric variable

formula:

$$g_\Psi = \frac{\Psi}{\sqrt{MS_{error}}} \qquad \text{(eq.32)}$$

where ψ is a contrast, and MS_{error} is the pooled within-groups mean squared error. Thus, g_ψ is a standardized mean difference for a contrast.

intuition: g_ψ could be viewed as an extension of Hedges's g to more than two groups: the denominator takes into account the variability of the data in all groups of the data set, regardless of whether they are featured in the contrast. g_ψ allows for a focused comparison of e.g. two groups, and in addition to this for focused comparisons of e.g. one group with a weighted ensemble of two other groups.

complements: analysis of variance & post-hoc tests

range: -inf to +inf; no effect: 0

confidence intervals: bootstrap (independent & dependent); approximate analytical (dependent); exact analytical (independent)

more information: (Kline, 2004)

examples: 1. Do the control and first treatment group differ in performance (post-treatment error of commission)?
mes1way(com_post,'g_psi','group',group,'cWeight',[1 -1 0]) yields g_ψ=-1.49. Exact analytical confidence intervals based on noncentral t distribution are [-2.16 -0.80], demonstrating a robust effect. The negative sign of g_ψ indicates that the average values in the treatment group are larger than those in the control group. Reversing the sign of the contrast weights results in identical numerical results with reversed sign.
2. Does the control group differ from the ensemble of both treatment groups (equally weighted) in terms of performance?
mes1way(com_post,'g_psi','group',group,'cWeight',[1 -0.5 -0.5]) yields g_ψ=-1.06 (CI95 [-1.63 -0.49]), attesting a strong effect to the treatment in both groups. The fact that now the value of g_ψ is lower than in the first comparison implies that the effect in the second treatment group is somewhat weaker than in the first treatment group.
3. The same question as in example 2 above, but now assuming dependent samples: g_ψ=-1.06 as above, because the definition of g_ψ is identical in the dependent and independent case. However, the confidence intervals differ; they are now [-1.46 -0.66]. These confidence intervals, based on the standard deviation of the difference scores, are approximate. They are slightly more narrow than under the assumption of independence above, which is to be expected as correlations between subjects were built into the data set.
Please note that the output of *mes1way.m* (**stats.g_psi**) will be a column array in which the first element is always NaN: by definition, the first position in all output arrays is reserved for the omnibus effect, which for g_ψ does not exist. From the second row on, **stats.g_psi** will hold the values for the contrasts (as many rows as contrasts are specified). The same applies to **stats.g_psiCi**, the confidence intervals.

comments: like Hedges's g it assumes homogeneity of variance in all groups

%%

name: **psibysd ($\psi/S_{D\psi}$)**

data structure: two or more dependent samples, metric variable

formula:

$$\frac{\Psi}{S_{D_\Psi}} \qquad \text{(eq.33)}$$

where ψ is a contrast, and $S_{D\psi}$ is the standard deviation of the difference scores of the contrast.

intuition: This MES is an extension of *mdbysd* to more than two groups. It differs from g_ψ in its standardizer (denominator), which is only defined for dependent (repeated measures) data. In dependent data sets with much variability between subjects this MES is more sensitive than g_ψ.

complements: repeated measures analysis of variance & post-hoc tests

range: -inf to +inf; no effect: 0

confidence intervals: bootstrap, approximate analytical

more information: (Kline, 2004)

examples: 1. Do the control and second treatment groups differ in performance (post-treatment error of commission)?
Given the information that the individuals in both groups were identical, we may wish to compute **mes1way(com_post,'psibysd','group',group,'cWeight',[1 0 -1],'isDep',1)**, which yields $\psi/S_{D\psi}$=-1.86. Approximate analytical confidence intervals are [-2.33 -1.39]. It is important to reiterate (Kline 2004) that this value cannot be compared quantitatively to g_ψ for the same data (which yields -0.63 [-0.79 -0.47]) or g_ψ from any other data set. The reason is that both MES are defined in different metrics: g_ψ in terms of the pooled standard deviation of the data set, $\psi/S_{D\psi}$ in terms of the standard deviation of the contrast's difference scores.
Note that as with g_ψ the output of mes1way.m (stats.psibysd) will be a column array in which the first element is always NaN.

comments: none

%%

name: **eta squared (η^2)**

data structure: two or more independent or dependent samples, metric variable

formula:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} \qquad \text{(eq.34)}$$

where SS_{effect} is the sum of squares between groups (treatments), and SS_{total} is the overall sum of squares, composed of SS_{effect} as defined above and SS_{error}, the sums of squares for the effect ANOVA error term (i.e. the within-groups sum of squares, SS_{within}). For contrasts, the numerator changes to SS_Ψ , the sum of squares of the contrast (the denominator remains the same).

intuition: how much variance in the metric variable is explained by group membership? ratio of the variance between groups and the total variance

complements: analysis of variance

range: 0 to +1; no effect: 0

confidence intervals: bootstrap (independent & dependent); exact analytical (independent)

more information: (Kline, 2004)

examples: 1. Do the three treatment groups differ in performance?
[p,tbl]=anova1(com_post,group) yields a significant main effect of group with F(2,57)=11.13 and p=0.00008. So, how much variance is explained by group differences?
mes1way(com_post,'eta2','group',group) yields η^2=0.28, tantamount to the statement that 28 % of the total variance in the dependent variable is accounted for by treatment. The relatively wide range of CI95, computed analytically via noncentral F distribution ([0.09 0.43]), highlights the variability in the data. Bootstrapped CI95, obtained via **mes1way(com_post,'eta2','group',group,'nBoot',10000)**, are [0.19 0.54].
2. What is the outcome of above question assuming dependent (repeated measures) data? **mes1way(com_post,'eta2','group',group,'isDep',1)** yields exactly the same numerical results: the definition of η^2 is identical for dependent and independent data (in contrast to partial η^2 and partial ω^2, see further below).
3. η^2 may also be computed for contrasts. In order to obtain values for two contrasts in one step, compute
mes1way(com_post,'eta2','group',group,'cWeight',[1 -1 0; 1 0 -1]).
η^2 values for the two contrasts are 0.28 and 0.05, respectively, confirming that the treatment effects are much more consistent in the first group. Exact analytical CI95 of η^2 for contrasts cannot be computed because the concrete formula, $SS_\Psi/(SS_{between} + SS_{total})$, is composed of three different terms and thus does not directly relate to F distributions. However, CI95 of partial η^2 are amenable to exact analytical computation, see there.

comments: frequently used

%%%

name: **partial eta squared (η_p^2)**

data structure: two or more independent or dependent samples, metric variable

formula: the general form of η_p^2 is

$$\eta_p^{\,2} = \frac{SS_{effect}}{SS_{effect} + SS_{error}} \qquad \text{(eq.35)}$$

where SS_{effect} is the sum of squares between groups (treatments) and SS_{error} is the sums of squares for the effect ANOVA error term (i.e. the within-groups sum of squares, SS_{within}). For contrasts, SS_{effect} corresponds to SS_Ψ , the sum of squares of the contrast:

$$\eta_p^{\,2} = \frac{SS_\psi}{SS_\psi + SS_{error}} \qquad \text{(eq.36)}$$

If the data are dependent (repeated measures) SS_{error} corresponds to $SS_{within\ x\ subject}$, the within-groups sum of squares minus the between-subjects sum of squares:

$$\eta_p^{\,2} = \frac{SS_{effect}}{SS_{effect} + SS_{within} - SS_{subject}} \qquad \text{(eq.37)}$$

intuition: how much variance in the metric variable is explained by group membership, corrected for variability between subjects?

complements: (repeated measures) analysis of variance

range: 0 to +1; no effect: 0

confidence intervals: bootstrap (independent & dependent); exact analytical (independent)

more information: (Kline, 2004)

example: 1. Do the three treatment groups differ in performance? **mes1way(com_post,'partialeta2','group',group,'cWeight',[1 -1 0; 1 0 -1])** yields η_p^2=0.28 [0.09 0.43] for the omnibus effect, the same value as for η^2 above. For the omnibus effect in a 1-way design with independent data partial eta squared is always identical to eta squared. However, the values for one of the two contrasts are different: 0.28 [0.10 0.44] and 0.07 [NaN 0.21], respectively. η_p^2 values computed for contrasts are always equal to or larger than the corresponding values for η^2 as the denominator is by design devoid of between-conditions sources of variability stemming from effects other than the one under consideration. Compare equations 30 and 32: while for η^2 the denominator contains $SS_{effect} + SS_{error} = SS_{between} + SS_{error}$, for η_p^2 it is $SS_\psi + SS_{error}$. Summed squares from groups not featuring in the contrast have been partialled out. Note also that the lower CI of the second contrast is too close to zero to be computable via the iterative procedure involving noncentral F distributions, hence its value is automatically set to NaN to indicate this fact.
2. Assuming that the subjects tested were identical in all three treatment groups, do the groups differ in performance? **mes1way(com_post,'partialeta2','group',group,'isDep',1)** yields η_p^2 =0.39, higher than under the assumption of an independent design. The reason is that in dependent 1-way designs 'subject' is a factor which can be partialled out, leading to smaller denominator values (eq. 37). Confidence intervals can be obtained via bootstrapping; including the same contrasts as used above for η^2 via
mes1way(com_post,'partialeta2','group',group,'cWeight',[1 -1 0; 1 0 -1], 'isDep',1,'nBoot',10000)
yields 0.39 [0.29 0.74], 0.39 [0.28 0.74] and 0.10 [0.04 0.43] for the main effect and the two contrasts, respectively.

comments: partial eta squared is always equal to or higher than eta squared
acknowledges that the total variance in the data is brought about by several independent variables, and that, hence, the more variables in the ANOVA, the smaller the contribution of each individual variable to the total variance

%%

name: **omega squared (ϖ^2)**

data structure: two or more independent or dependent samples, metric variable

formula: For independent data, $\omega^2 = \frac{SS_{effect} - (J-1)\,MS_{error}}{SS_{total} + MS_{error}}$ (eq.38)

where SS_{effect} is the sum of squares between groups (treatments), SS_{total} is the overall sum of squares, J is the number of levels of the factor (groups), and MS_{error} is the mean squared error within groups.
For dependent data assuming an additive model (i.e. no subject x treatment interaction),

$$\omega^2 = \frac{\hat{\sigma}^2_{effect}}{\hat{\sigma}^2_{effect} + \hat{\sigma}^2_{subject} + \hat{\sigma}^2_{error}} \quad \text{(eq.39)}$$

where

$$\hat{\sigma}^2_{effect} = \frac{J-1}{JN}(MS_{effect} - MS_{withinxsubject})$$

$$\hat{\sigma}^2_{subject} = \frac{1}{J}(MS_{subject} - MS_{withinxsubject})$$

$$\hat{\sigma}^2_{error} = MS_{withinxsubjectj}$$

and (in addition to the notation above) N is the number of samples per group, MS_{effect} is the between-groups mean squares, $MS_{within\ x\ subject}$ is the within-groups x subjects mean squares, and $MS_{subject}$ the between-subjects mean squares. For contrasts,

$$\hat{\sigma}^2_{effect} = \frac{1}{JN}(SS_\psi - MS_{withinxsubject})$$

where SS_Ψ is the mean summed squares for the contrast. Note that while η^2 is a positively biased descriptive measure of effect size, ω^2 is an inferential measure which corrects for the bias due to small sample sizes, hence the different symbols.

intuition: how much variance in the metric variable is explained by group membership? ratio of the variance between groups and the total variance

complements: analysis of variance

range: 0 to +1; no effect: 0

confidence intervals: bootstrap (independent & dependent); exact analytical (independent)

more information: (Kline, 2004)

examples: 1. Do the three treatment groups differ in performance?
mes1way(com_post,'omega2','group',group) yields ω^2=0.25 [0.06 0.41], possibly a more realistic estimate of variance explained by group membership given the comparatively low sample size (20 per group). Taking into account a repeated measures design and computing CI95 via bootstrapping yields ω^2=0.25 [0.17 0.50]. ω^2 for contrasts may be computed in the same manner as explained for η^2.
2. Do the treatment groups differ in their *pre*drug performance?
[p,tbl] = anova1(com_pre,group) yields a nonsignificant main effect of group with F(2,57)=0.68 and p=0.51. Still, one can ask how much variance is explained by group: **mes1way(com_pre,'omega2','group',group)** yields ω^2=-0.01. Negative values may result when MS_{error}, the mean squared error within groups ($MS_{within\ x\ subject}$ for dependent data) is large compared to $SS_{between}$ so that the numerator in the formulae above attains a negative value. As Kline (Kline, 2004) noted, weak effects or small samples are the likely cause. Accordingly, the lower bootstrapped CI95 is 0. The upper CI95 of 0.09 confirms that the effect is likely very small.

comments: less frequently used than eta squared; has the advantage of being a more accurate estimator of the population effect size at small sample sizes, but the disadvantage of being less intuitive

%%
name: **partial omega squared ($\varpi_p{}^2$)**

data structure: two or more independent or dependent samples, metric variable

formula: For independent data,

$$\omega_p{}^2 = \frac{SS_{effect} - (J-1)\,MS_{error}}{SS_{effect} + (N-J+1)\,MS_{error}} \qquad \text{(eq.40)}$$

where SS_{effect} is the sum of squares between groups (treatments), J is the number of levels of the factor (groups), and MS_{error} is the mean squared error within groups.
For dependent data (assuming an additive model as for partial eta squared),

$$\omega_p{}^2 = \frac{\hat{\sigma}^2_{effect}}{\hat{\sigma}^2_{effect} + \hat{\sigma}^2_{error}} \qquad \text{(eq.41)}$$

where

$$\hat{\sigma}^2_{effect} = \frac{J-1}{JN}(MS_{effect} - MS_{withinxsubject})$$

$$\hat{\sigma}^2_{error} = MS_{withinxsubjectj}$$

and (in addition to the notation above) N is the number of samples per group, MS_{effect} is the between-groups mean squares, $MS_{within\ x\ subject}$ is the within-groups x subjects mean squares, and $MS_{subject}$ the between-subjects mean squares. For contrasts,

$$\hat{\sigma}^2_{effect} = \frac{1}{JN}(SS_{\psi} - MS_{withinxsubject})$$

where SS_{ψ} is the mean summed squares for the contrast.

intuition: how much variance in the metric variable is explained by group membership? ratio of the variance between groups and the total variance

complements: analysis of variance

range: 0 to +1; no effect: 0

confidence intervals: bootstrap (independent & dependent); exact analytical (independent)

more information: (Kline, 2004)

examples: 1. Do the three treatment groups differ in performance?
mes1way(com_post,'partialomega2','group',group) yields ω_p^2=0.25 [0.06 0.41], the same values as for ω^2. The same reasoning as in the case of η^2 and η_p^2 applies: in 1-way designs with independent samples the values are identical for omnibus comparisons..
2. Taking into account a repeated measures design and computing CI95 via bootstrapping yields ω^2=0.27 [0.18 0.64]. ω^2 for contrasts may be computed in the same manner as explained for η^2.

comments: like partial eta squared, partial omega squared disregards non-effect between-conditions variance sources; therefore, its value for contrasts or in analyses with dependent data always has a larger value than omega squared. Like omega squared, it is a more accurate estimator of the population effect size at small sample sizes.

Effect size measures covered in mes2way.m

Introductory notes
There are numerous ways of conducting factorial analyses with two or more factors. The model can include or exclude interactions between the factors, factors may be fixed or random, the data may be dependent across one or both factors; in case of independent data cell sizes may be balanced or unbalanced, and in the latter case summed squares may be computed in several different ways. Therefore, we restricted the defaults and set of options to what we believe is a reasonable standard and starting point for further developments. Concretely,

- factors are fixed
- interactions between the factors are included in the underlying model
- data may be dependent across either of the factors (mixed between-subjects design), both (completely within-subjects), or none (completely between-subjects)
- the data may be unbalanced along the between-subjects factor(s) (if any)
- 'Method 1/Type III' summed squares are computed (which is also the standard in the Matlab ANOVA functions). This implies that if the data is unbalanced the cause is an occasional, random loss of a datum in an experimental design that strives for equal cell sizes

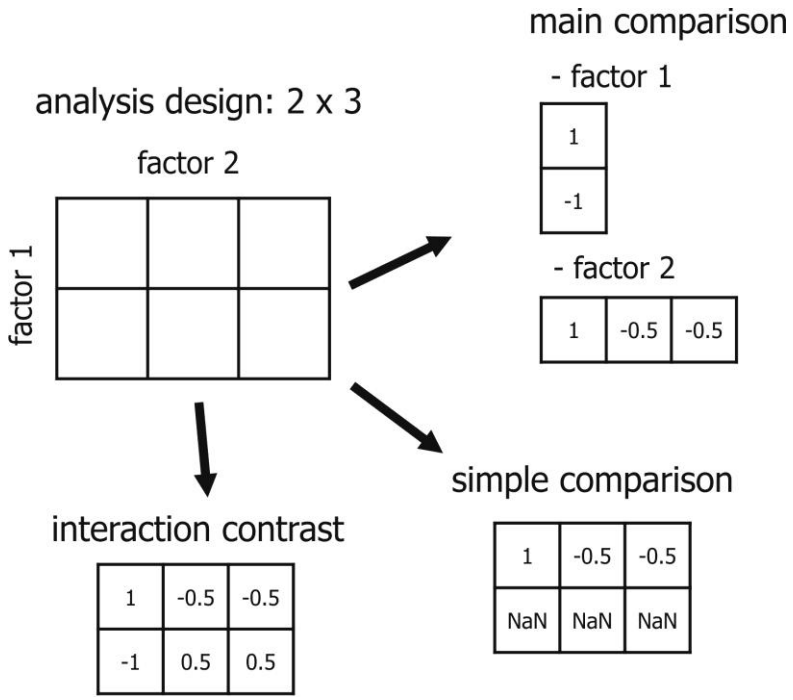

Figure 8. Example of a 2-way analysis design, different types of contrast analysis and required format of the input variable cWeight. The large grid in the upper left part represents a 2 by 3 factorial design. In main comparisons (upper right), contrasts of marginal means are computed; accordingly, cWeight must be a 2-element column vector or a 3-element row vector. In the simple comparison shown in the lower right corner, a contrast of the factor 2 effect at level 1 of factor 1 is computed. Note that cWeight must be of the same dimensions as the analysis design and that except for the row or column of interest the entries must be NaN. The lower left corner represents contrast weights for an interaction contrasts.

Like its 1-way counterpart, *mes2way.m* computes contrast-related measures of effect size. 2-way contrasts fall into two categories: single factor contrasts and interaction contrasts. Among the former, there may be simple or main comparisons. Input variable 'cWeight' (in short, c) must be shaped accordingly (Figure 8):

- MAIN COMPARISON CONTRAST: c must be a single-column array (comparison of levels of the first factor) or a single-row array (ditto for second factor)
- SIMPLE COMPARISON CONTRAST: c must match the analysis design, i.e. in a 2x3 analysis c must have two rows and three columns, but all elements except the row or the column of interest must be NaN
- INTERACTION CONTRAST: c must match the analysis design and it should also be doubly centered, that is, row sums and column sums must all be zero

Please note that mes2way.m, in contrast to its 1-way counterpart, allows only one set of contrast weights to be specified in its optional input variable 'cWeight'.

The **example computations** are 2-way analyses with group (experimental treatment) the first factor and gender the second factor. In some of the terms listed below subscripts for the two factors are A and B, that is, e.g. MS_A are the mean squares of the first factor, MS_B those of the second factor, and MS_{AB} those of the interaction of both.

Finally, we wish to highlight the option to **plot the data and illustrate the layout of the analysis**, including contrast weights. For example, before we do any calculation of MES we may inspect the example data, now cast as a twoway analysis as explained above:

mes2way(com_post,[group sex],{},'doDataPlot',1);

will produce a plot as shown in Figure 9, left. The three subplot rows correspond to the three experimental groups and the columns correspond to gender. Attributing tags to the factors and assuming dependence of the data along the first factor via

mes2way(com_post,[group sex],{},'fName',{'group', 'sex'},'isDep',[1 0],'doDataPlot',1);

results in Figure 9 center, in which the colors of the data points identify individuals. Finally, a plot of an exemplary contrast-based computation like

mes2way(com_post,[group sex],{},'fName',{'group', 'sex'},'isDep',[1 0],...
'cWeight',[1 NaN;-1 NaN;0 NaN],'doDataPlot',1);

is shown in Figure 9 right, in which a red background corresponds to a positive contrast weight (blue=negative weight, gray=zero contrast weight, white=not featured in contrast).

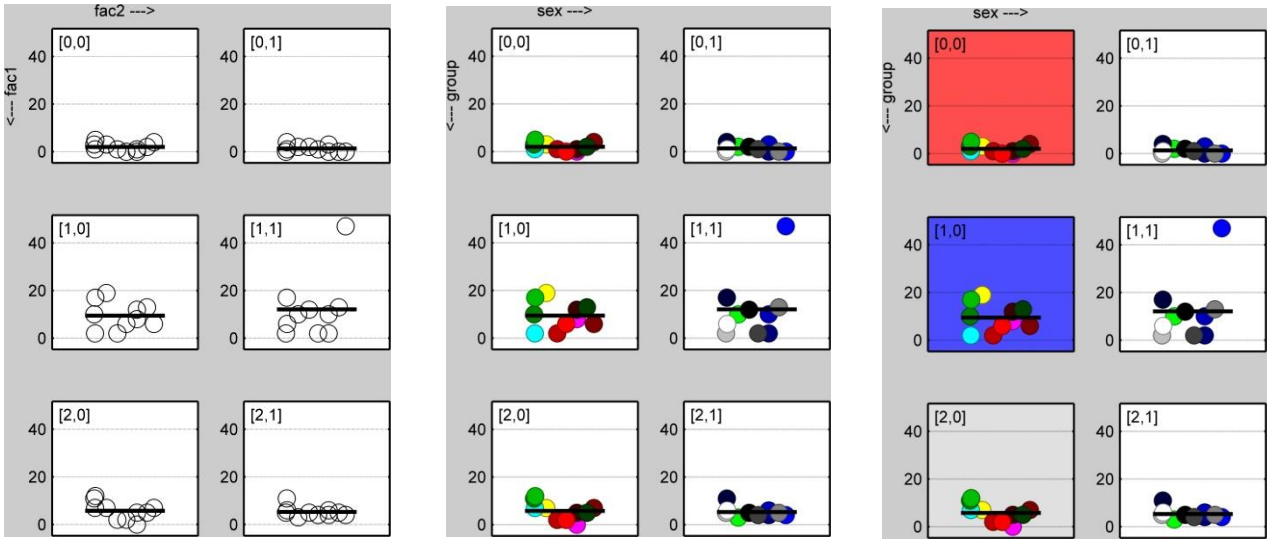

Figure 9. Examplary plots of the example data cast as a twoway analysis, produced by mes2way.m. See main text for details.

%%

name: **psi (ψ)**

data structure: factorial 2-way design, metric variable

formula:

$$\Psi = \sum_{i=1}^{a} c_i m_i \quad \text{(main comparison of factor A)} \qquad \text{(eq.42)}$$

where *a* is the number of levels of factor A, m_i is the i^th marginal mean of the factor (weighted by sample size), and c_i is the corresponding contrast weight;

$$\Psi = \sum_{i=1}^{a} c_i m_i \quad \text{(simple comparison of factor A at specific level of factor B)} \qquad \text{(eq.43)}$$

where *a* is the number of levels of factor A, m_i the i^th mean of groups of factor A at the chosen level of factor B, and c_i is the corresponding contrast weight;

$$\Psi = \sum_{i=1}^{a} \sum_{j=1}^{b} c_{ij} m_{ij} \quad \text{(interaction contrast)} \qquad \text{(eq.44)}$$

where *a* and *b* are the numbers of levels of factors A and B, respectively, m_{ij} is the group mean at level *i* of factor A and level *j* of factor B, and c_{ij} is the corresponding contrast weight

intuition: twoway equivalent of mean differences, expressed in the original units of measurement

complements: analysis of variance & post-hoc tests

range: -inf to +inf; no effect: 0

confidence intervals: exact analytical (all designs), bootstrap (all designs). Analytical confidence intervals for designs with dependent data across one or both factors are derived from estimates of the standard error of the contrast based on standard ANOVA terms. See the introductory notes to mes1way.

more information: Kline (2004), Rosenthal et al. (2000)

examples: 1. Do the control and first treatment group differ in performance? This question may be cast in two different ways. First, we may perform a main comparison along the first factor: **mes2way(com_post,[group sex],'psi', 'cWeight',[1 -1 0]')** yields ψ=-9.15. Exact analytical confidence intervals are [-13.12 -5.18]. Compare this to the results of the corresponding 1-way analysis: the values are very similar, as is expected (and desirable) for an analysis designed to ignore the effect of the second factor. Second, we may perform simple comparisons along the first factor (treatment) for different levels of the second factor (gender): **mes2way(com_post,[group sex],'psi','cWeight',[1 NaN;-1 NaN;0 NaN])** yields ψ=-7.50. Exact analytical confidence intervals are [-13.12 -1.88]. In words, men in the first treatment group commit 7.5 errors more than those in the control group (placebo treatment). However, the uncertainty of this result as expressed in the CI95 is much larger than in the main comparison. This is to be expected given that the sample size of the simple comparison is half of that of the main comparison. For women, ψ=-10.80 [-16.42 -5.18].

2. The previous analysis demonstrated that the first treatment clearly has a sizeable effect. Ignoring this insight, we may pursue the (implausible) idea that the second treatment has opposite effects in men and women: it could increase commission error rates in men and decrease them in women. In other words, this would be a very specific *interaction* of treatment with gender. Standard analysis of variance will deliver F and p values for this interaction (e.g. via **mes2way(com_post,[group sex],{})**), but is generally not very helpful because i) F and p do not inform us on effect size, and ii) our interest lies in an interaction which is more specific than the one addressed by the ANOVA term. The problem can be cast as an interaction contrast via **mes2way(com_post,[group**

sex],'psi','cWeight',[1 -1;0 0;-1 1]). In words, we first compute the second treatment's effect for men and women separately, namely as the number of errors in the control group minus the number of errors in the second treatment group, and then subtract these results from each other. The result is ψ=0.20, a gender-specific difference of the treatment of much less than a single error. This difference is small, particularly compared to the variability in the number of commission errors in the whole data set (the standard deviation of commission errors pooled over all groups is 6.3). However, the confidence intervals, [-7.75 8.15], are quite wide. In order to enlarge our confidence in the result of a virtually absent interaction of second treatment with gender we would have to increase sample sizes.

comments: see the introductory paragraph on contrasts in this document. Also, once again it shall be stated that in some applications the real units attached to unstandardized statistics may convey a better idea of the difference between groups, the drawback being that the numerical values cannot readily be compared to unstandardized differences derived in other contexts. See g_psi below.

%%

name: **g_psi (g_ψ)**

data structure: factorial 2-way design, metric variable

formula:

$$g_\Psi = \frac{\Psi}{\sqrt{MS_{error}}} \qquad \text{(eq.45)}$$

where ψ is a contrast, and MS_{error} is the pooled within-cell mean square. Thus, g_ψ is a standardized mean difference for a contrast.

intuition: g_ψ could be viewed as an extension of Hedges's g to a 2-way factorial design: the denominator takes into account the variability of the data in all cells of the data set, regardless of whether they are featured in the contrast (see Kline (2004) for other standardizers). g_ψ allows for a focused comparison of specific subsets of the data.

complements: analysis of variance & post-hoc tests

range: -inf to +inf; no effect: 0

confidence intervals: exact analytical (completely between-subjects), bootstrap (all other designs)

more information: (Kline, 2004)

examples: 1. Do the control and first treatment group differ in performance? This question may be cast in two different ways. First, we may perform a main comparison along the first factor: **mes2way(com_post,[group sex],'g_psi', 'cWeight',[1 -1 0]')** yields g_ψ=-1.46. Exact analytical confidence intervals based on noncentral t distribution are [-2.13 -0.78]. Compare this to the results of the corresponding 1-way analysis: the values are very similar, as is expected (and desirable) for an analysis designed to ignore the effect of the second factor. Second, we may perform simple comparisons along the first factor (treatment) for different levels of the second factor (gender): **mes2way(com_post,[group sex],'g_psi','cWeight',[1 NaN;-1 NaN;0 NaN])** yields g_ψ=-1.20. Exact analytical confidence intervals are [-2.1 -0.29]. In words, the mean performance of men after the first treatment is 1.2 'standard deviations' (the square root of the pooled within-cell mean squares) away from that after the placebo treatment. However, the uncertainty of this result as expressed in the CI95 is much larger than in the main comparison. This is to be expected given that the sample size of the simple comparison is half of that of the main comparison. For women, g_ψ=-1.72 [-2.65 -0.78].

comments: like ~~Hedges' g~~Hedges's g it assumes homogeneity of variance in all groups.

%%%
name: **eta squared (η^2)**

data structure: factorial 2-way design, metric variable

formula: the general form is $\eta^2 = \frac{SS_{effect}}{SS_{total}}$ (eq.46)

where SS_{effect} is the effect sum of squares and SS_{total} is the overall sum of squares. Concretely, for the main effect of the first factor, $SS_{effect} = SS_A$, for the main effect of the second factor, $SS_{effect}=SS_B$, for the interaction effect, $SS_{effect}=SS_{AB}$, and for contrasts, $SS_{effect} = SS_\Psi$ .The denominator is $SS_A+SS_B+SS_{AB}+SS_{error}$.

intuition: how much variance in the metric variable is explained by group membership? ratio of the variance between groups and the total variance

complements: analysis of variance

range: 0 to +1; no effect: 0

confidence intervals: bootstrap

more information: (Kline, 2004)

examples: 1. Is there a difference in performance across treatment groups and gender? **mes2way(com_post,[group sex],'eta2')** computes η^2 values for both main effects and the interaction effect. For the first factor (treatment) η^2=0.28, as in the 1-way case. η^2 of the second factor is almost zero, which means that gender contributes very little (< 1%) to the total variability of the data. Confidence intervals of η^2 cannot be computed analytically via noncentral F distribution. Bootstrapped CI95 values are [0.19 0.54] and [0 0.06] for treatment and gender, respectively. η^2 values are identical for all designs (between vs. within subjects-factors).
2. η^2 may also be computed for contrasts: **mes2way(com_post,[group sex],'eta2','cWeight',[1 -1 0]')** yields η^2=0.28, barely smaller than the main treatment effect (a difference which is visible only in the third decimal digit).

comments: frequently used

%%%
name: **partial eta squared (η_p^2)**

data structure: factorial 2-way design, metric variable

formula: the general form of η_p^2 is

$$\eta_p^{\ 2} = \frac{SS_{effect}}{SS_{effect} + SS_{error}} \quad \text{(eq.47)}$$

where SS_{effect} is the effect sum of squares and SS_{error} is the sum of squares for the effect ANOVA error term. In analogy to the 1-way case, the concrete terms depend on the design of the analysis; in 2-way designs they also depend on which effect (main, interaction, contrast-related) is under consideration. As an example, for the effect of the first factor (treatment) in a mixed within-subjects analysis with treatment being the within-subjects factor,

$$\eta_p^{\ 2} = \frac{SS_A}{SS_A + SS_{Axsubj/B}} \quad \text{(eq.48)}$$

where $SS_{A\ x\ subj/B}$ is the within-subjects sum of squares, computed separately for each level of the between-subjects factor and added up. This term excludes variability due to the between-subjects factor and the individuals (subjects) themselves.

intuition: how much variance in the metric variable is explained by group membership, corrected for sources of variability not to do with the effect(s) under consideration?

complements: (repeated measures) analysis of variance

range: 0 to +1; no effect: 0

confidence intervals: exact analytical or bootstrap (completely between-subjects)

more information: (Kline, 2004)

examples: 1. Is there a difference in performance across treatment groups and gender? **mes2way(com_post,[group sex],'partialeta2')** yields η_p^2 =0.28 [0.09 0.44] (exact analytical CI95) for the first factor, treatment. This is surprising insofar as η_p^2 values are generally expected to be larger than η^2 values. Examination of the ANOVA table of results in the command window reveals the reason: the summed squares of the second factor and those of the interaction between both factors are dwarfed by the summed squares of the first factor and SS_{within}, the within-cells sums of squares. As the former two terms are all the difference between the denominators in the partialled and unpartialled variant of eta squared, their presence or absence is negligible for the whole expression. However, the picture changes if treatment is assumed to be a within-subjects factor. Now, η_p^2 =0.40, illustrating that interindividual variability masked a substantial proportion of the treatments' effects, similar to the insight gleaned from the 1-way analysis.

%%

name: **omega squared (ϖ^2)**

data structure: factorial 2-way design, metric variable

formula: the general form is $\omega^2 = \frac{\hat{\sigma}^2_{effect}}{\hat{\sigma}^2_{total}}$ (eq.49)

As in the case of η^2 the terms depend on the design (presence of within-subjects factors) and on the factor(s) under consideration. Kline (Kline, 2004) presents tables of variance component estimators for various analysis designs; these can be used to assemble the concrete formulas. For example, the (simplified) formula for a completely between-subjects design is

$$\omega^2 = \frac{SS_{effect} - df_{effect}\, MS_{error}}{SS_{total} + MS_{error}} \quad \text{(eq.50)}$$

where df_{effect} is the degrees of freedom of the factor(s) under consideration, and all other terms are the same as explained before

intuition: how much variance in the metric variable is explained by group membership? ratio of the unbiased estimated variance between groups and the unbiased estimated total variance

complements: analysis of variance

range: 0 to +1; no effect: 0

confidence intervals: bootstrap

more information: (Kline, 2004)

examples: 1. Do the three treatment groups differ in performance? **mes2way(com_post,[group sex],'omega2','nBoot',10000)** yields, for factor treatment, ω^2=0.25 [0.16 0.52], similar to the 1-way equivalent. Please note that *mes2way.m* does not compute ω^2 for data with within-subjects factors.

%%%

name: **partial omega squared (ϖ_p^2)**

data structure: factorial 2-way design, metric variable

formula: the general form is $$\omega_p^{\;2} = \frac{\hat{\sigma}^2_{effect}}{\hat{\sigma}^2_{effect} + \hat{\sigma}^2_{error}}$$ (eq.51)

As for η^2 and ω^2 illustrated above, the terms depend on the design (presence of within-subjects factors) and on the factor(s) under consideration. Kline (2004) presents tables of variance component estimators for various analysis designs; these can be used to assemble the concrete formulas. For example, the (simplified) formula for a completely between-subjects design is

$$\omega_p^{\;2} = \frac{df_{effect}(F_{effect} - 1)}{df_{effect}(F_{effect} - 1) + N}$$ (eq.52)

where df_{effect} and F_{effect} are the degrees of freedom and F value of the factor(s) under consideration, respectively, and N is the total number of samples.

intuition: how much variance in the metric variable is explained by group membership? ratio of the variance between groups and the total variance

complements: analysis of variance

range: 0 to +1; no effect: 0

confidence intervals: exact analytical or bootstrap (completely between-subjects)

more information: (Kline, 2004)

examples: 1. Do the three treatment groups differ in performance? **mes2way(com_post,[group sex],'partialomega2')** yields, for factor treatment, ω_p^2=0.25 [0.05 0.41], similar to the 1-way equivalent. Please note that *mes2way.m* does not compute ω_p^2 for data with within-subjects factors.

Effect size measures covered in mestab.m

%%

name: **risk difference (RD)**

data structure: absolute frequency in a fourfold table

formula: $\pi_C - \pi_T$ (eq.53)

where π_C is proportion of relapses in the control group and π_T is the proportion of relapses in a treatment group

intuition: difference between proportion relapsed in control group and proportion relapsed in treatment group

complements: chi square, Fisher's exact test

range: –1 to +1; no effect: 0; +1 indicates maximum superiority of treatment, –1 indicates maximum adverse effect of treatment

confidence intervals: analytical

more information: (Kline, 2004)

example: smoking and coronary heart disease (CHD): 78 of 137 smokers have CHD, so the risk for this group is 78/137=0.57. 42 out of 103 nonsmokers have CHD, too, so the risk is 42/103=0.41. The observed risk difference is *RD*=0.57–0.41=0.16, so smokers have a 16% greater risk for CHD than nonsmokers (CI95 [0.04 0.29]. Code: **mestab([smokesick,smokewell;nsmokesick,nsmokewell])**

comments: also known as "proportion difference"

%%

name: **risk ratio (RR)**

data structure: absolute frequency in a fourfold table

formula: $\frac{\pi_C}{\pi_T}$ (eq.54)

where π_C is proportion of relapses in the control group and π_T is the proportion of relapses in a treatment group

intuition: ratio between proportion relapsed in control group and proportion relapsed in treatment group; how much more likely is a relapse in the control group? if RR<1, relapse in the control group is less likely

complements: chi square, Fisher's exact test

range: 0 to +inf; no effect: 1;
0 indicates no relapses in control group, a high number indicates superiority of treatment, –1 indicates maximum adverse effect of treatment

confidence intervals: analytical

more information: (Kline, 2004)

example: smoking and coronary heart disease (CHD): 78 of 137 smokers have CHD, so the risk for this group is 78/137=0.57. 42 out of 103 nonsmokers have CHD, too, so the risk is 42/103=0.41. The observed risk ratio is *RR*=0.57/0.41=1.4 with CI95 [1.06 1.84], so smokers have a 1.4 times higher risk for for CHD than nonsmokers; code: **mestab([smokesick,smokewell;nsmokesick,nsmokewell])**

comments: also known as “relative risk” or “rate ratio”

%%
name: **odds ratio (ϖ, OR)**

data structure: absolute frequency in a fourfold table

formula:

$$\varpi = \frac{\Omega_C}{\Omega_T} = \frac{\pi_C/(1-\pi_C)}{\pi_T/(1-\pi_T)} \quad \text{(eq.55)}$$

where π_C is proportion of relapses in the control group and π_T is the proportion of relapses in a treatment group; accordingly, Ω_C is the odds for relapse in the control population and Ω_T is the odds for relapse in the treatment population

intuition: ratio of the within-populations odds for the undesirable outcome

complements: chi square, Fisher’s exact test

range: 0 to +inf; no effect: 1;
0 indicates no relapses in control group, a high number indicates superiority of treatment, –1 indicates maximum adverse effect of treatment

confidence intervals: analytical

more information: (Kline, 2004)

example: smoking and coronary heart disease (CHD): 78 of 137 smokers have CHD, so the odds for this group are 78/59=1.32. 42 out of 103 nonsmokers have CHD, too, so the odds in this group are 42/61=0.69. The observed odds ratio is ϖ=1.32/0.69=1.92 with CI95 [1.14 3.22], so smokers have twice the odds for CHD compared to smokers. Code: **mestab([smokesick,smokewell;nsmokesick,nsmokewell])**

comments: ϖ has nothing to do with ϖ^2, the effect size used for analysis of variance

%%
name: **phi (φ)**

data structure: absolute frequency in a fourfold table

formula:

$$\varphi = \frac{\pi_{CR}\pi_{TNR} - \pi_{CNR}\pi_{TR}}{\sqrt{\pi_C\pi_T\pi_R\pi_{NR}}} \quad \text{(eq.56)}$$

where π_C is proportion of relapses in the control group, π_T is the proportion of relapses in a treatment group, π_R is the proportion of cases relapsed, π_{NR} is the proportion of cases not relapsed, π_{CR} is the proportion relapsed in the control group, π_{TR} is the proportion relapsed in the treatment group, π_{TCNR} is the proportion not relapsed in the control group, and π_{TNR} is the proportion not relapsed in the treatment group

intuition: Pearson correlation between the row and column variables; can be interpreted the same way

complements: chi square, Fisher's exact test

range: –1 to +1; no effect: 0; Cohen suggested φ=0.1, φ=0.3, and φ=0.5 to denote small, medium, and large effects, respectively

confidence intervals: exact analytical

more information: (Kline, 2004)

example: smoking and coronary heart disease (CHD)
The strength of the relationship between smoking and CHD, expressed as a correlation, is φ=0.16, with exact analytical CI95 [0.07 0.29].

comments: is algebraically equivalent to Pearson's product-moment correlation computed for two dichotomous variables (e.g. relapsed and not relapsed (coded as 1 and 0, respectively) for two groups)

%%

name: **sensitivity**

data structure: absolute frequency in a fourfold table

formula:

$$sensitivity = \frac{p(P \mid D)}{p(D)} \qquad \text{(eq.57)}$$

where *p(P|D)* is the proportion of cases diagnosed with a disease among those cases that actually suffer from that disease, divided by the total number of cases suffering from that disease, *p(D)*

intuition: probability of being positively diagnosed with a disease when actually suffering from that disease

complements: chi square, Fisher's exact test

range: 0 to +1; generally speaking: the higher, the better

confidence intervals: not applicable

more information: (Kline, 2004)

example: mammography screening: 1,000 women (age 40) are screened for breast cancer; 10 of them actually have breast cancer, 8 of them are screened positively by mammography; thus, the sensitivity of mammography is 8/10=0.8

comments: high sensitivity, by itself, is of little value when specificity is low

%%%

name: **specificity**

data structure: absolute frequency in a fourfold table

formula:

$$specificity = \frac{p(\neg P \mid \neg D)}{p(\neg D)} \quad \text{(eq.58)}$$

where *p(¬P|¬D)* is the proportion of cases not diagnosed with a disease among those cases that do not have that disease, divided by the total number of cases that do not have that disease, *p(¬D)*

intuition: probability of being not diagnosed with a disease when actually not having that disease

complements: chi square, Fisher's exact test

range: 0 to +1; generally speaking: the higher, the better

confidence intervals: not applicable

more information: (Kline, 2004)

example: mammography screening: 1,000 women (age 40) are screened for breast cancer; 990 of them are healty; still, 99 of them are (incorrectly) tested positive with mammography; thus, the specificity of mammography is 891/990=0.9 (which is quite good)

comments: high specificity, by itself, is of little value when sensitivity is low

%%%

name: **positive predictive value (PPP)**

data structure: absolute frequency in a fourfold table

formula:

$$positive\ predictive\ value = \frac{p(D \mid P)}{p(P)} \quad \text{(eq.59)}$$

where *p(D|P)* is the proportion of cases having a disease among all those cases tested positive for a disease, and *p(P)* is the proportion of cases tested positive for a disease

intuition: probability of having a disease when tested positive for that disease

complements: chi square, Fisher's exact test

range: 0 to +1; generally speaking: the higher, the better

confidence intervals: not applicable

more information: (Kline, 2004)

example: mammography screening: 1,000 women (age 40) are screened for breast cancer; 990 of them are healty; overall, 107 of them are tested positive with mammography; thus, the positive predictive value of mammography is 8/107=0.07 (which is quite low)

comments: none

%%%
name: **negative predictive value (NPP)**

data structure: absolute frequency in a fourfold table

formula:

$$positive\ predictive\ value = \frac{p(\neg D \mid \neg P)}{p(\neg P)} \quad \text{(eq.60)}$$

where *p(¬D|¬P)* is the proportion of cases not having a disease among all those cases tested negative for a disease, and *p(¬P)* is the proportion of cases tested negative for a disease, regardless of true status

intuition: probability of not having a disease when tested negative for that disease

complements: chi square, Fisher's exact test

range: 0 to +1; generally speaking: the higher, the better

confidence intervals: not applicable

more information: (Kline, 2004)

example: mammography screening: 1,000 women (age 40) are screened for breast cancer; 893 are tested negative; of these, 891 are truly healthy; thus, the negative predictive value of mammography is 891/893=0.99776

comments: none

%%%
name: **binomial effect size display (BESD)**

data structure: absolute frequency in a fourfold table

formula:

$$success_{experimental} = .50 - r/2$$
$$success_{control} = .5 + r/2$$

where r corresponds to the phi coefficient (Pearson correlation between two dichotomous variables), success
the BESD is created by setting raw frequencies in a fourfold table into standardized percentages, such that all margins are equal (see Table 4 for an example)

intuition: change in success rate of (e.g. survival or improvement of a condition) brought about by a new treatment procedure

complements: chi square, Fisher's exact test, phi correlation

range: 0 to +1; generally speaking: the higher, the better

confidence intervals: not applicable

more information: (Rosenthal and Rubin, 1981)

example: Rosenthal and Rubin (1981) give an example where the correlation between two dichotomous measures (treatment vs. control, alive vs. dead) was a mere .32, and point out that this translates to an increase in survival rate from 34% to 66%, brought about by the treatment

comments: Rosenthal and Rubin (1981) suggest using the BESD also when the two variables are not dichotomous but metric
although the authors label the BESD a "simple" display, that is "easily understood by researchers, students, and lay persons", the interpretation of the BESD is all but straightforward; see e.g. (Randolph and Edmondson, 2005)

	Treatment outcome		
Condition	**Alive**	**Dead**	Σ
Treatment	66	34	100
Control	34	66	100
Σ	100	100	200

Table 4: The Binomial Effect Size Display for a correlation of phi=0.32, accounting for 10% of the variance (from Rostenthal & Rubin, 1981); this correlation translates to an increase in the survival rate from 34% to 66% (i.e., reducing the death rate from 66% to 34%).

%%

name: **Cramer's V**

data structure: absolute frequency in rectangular tables; can have more than 2 rows and 2 columns

formula:

$$V = \sqrt{\frac{\chi^2(r-1, c-1)}{\min(r-1, c-1) \times N}} \quad \text{(eq.61)}$$

where χ^2 is the chi-square statistic with degrees of freedom equal to the product of the number of rows (r) minus one and the number of columns (c) minus one; the denominator holds the product of the sample size, N, and the smaller dimension of the table minus one

intuition: can be interpreted as a correlation coefficient for 2*2 tables

complements: chi square, Fisher's exact test

range: 0 to +1, with 1 indicating the strongest possible relationship

confidence intervals: exact analytical

more information: (Kline, 2004)

example: taken from Kline, 2004 (pp157), based on Woodward et al. (2000), who examined the relation between attentional problems assessed at age 13 and driving records assessed at age 21; portions of the data is presented in Table 5.
An increase in the percentage of cases caught driving without license is readily visible in the raw data and highly significant (p<0.001). The strength of the relationship, however, is moderate at best, with C*ramer's V* = 0.23, which results from **mestab([39 19 44 13 15; 436 166 144 34 31])**. Exact CI95 are [0.17 0.30].

comments: extension of the phi coefficient to larger tables; if r=c=2, V reduces to phi

Driving outcome at age 21	Attentional difficulties at age 13						
	1 (low)	2	3	4	5 (high)		
n	475	185	188	47	46	X^2	Cramer's V
Driving without license (%)	8.2	10.3	23.4	27.7	32.6	50.21	.23

Table 5 Data and results of example calculation for Cramer's V

Reference List

Bamber D (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology 12:387-415.

Bonett DG (2009) Meta-analytic interval estimation for standardized and unstandardized mean differences. Psychol. Methods 14:225–238.

Bonett DG (2015) Interval estimation of standardized mean differences in paired-samples designs. J. Educ. Behav. Stat. 20:1–11.

Borenstein M (2009) Effect sizes for continuous data. In The Handbook of Research Synthesis and Meta-Analysis, H. Cooper, L.V. Hedges, and J.C. Valentine, eds. (New York: Russel Sage Foundation), pp. 221–235.

Cohen J (1988) Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cumming G (2012) Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. (New York: Routledge).

Cureton EE (1956) Rank-biserial correlation. Psychometrika 21:287-290.

Cureton EE (1968) Rank-Biserial Correlation When Ties Are Present. Educational and Psychological Measurement 28:77-&.

Glass GV (1966) Note on Rank Biserial Correlation. Educational and Psychological Measurement 26:623-&.

Glass GV (1976) Primary, secondary, and meta-analysis of research. Educational Researcher 10:3-8.

Grissom RJ, Kim JJ (2012) Effect Sizes For Research: Univariate and Mutlivariate Applications (New York: Routledge).

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29-36.

Hedges LV (1981) Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics 6:107-128.

Hentschke H, Stüttgen MC (2011) Computation of measures of effect size for neuroscience data sets. European Journal of Neuroscience 34:1887-1894.

Howell DC (2002) Statistical Methods for Psychology. Thomson Learning.

Jordan D, Steiner M, Kochs EF, Schneider G (2010) A program for computing the predicition probability and the related receiver operating characteristic graph. Anesthesia & Analgesia 111:1416-1421.

Kline RB (2004) Beyond significance testing. Washington, DC: American Psychological Association.

Lakens D (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4:863.

MacMillan NA, Creelman CD (2005) Detection theory: a user's guide. New Jersey: Lawrence Erlbaum Associates, Inc.

McGraw KO, Wong SP (1992) A common language effect size statistic. Psychological Bulletin 111:361-365.

McNicol D (2005) A primer of signal detection theory. Mahwah, New Jersey: Lawrence Erlbaum Associates,Inc.

Nakagawa S, Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol. Rev 82:591–605.

Obuchowski NA, Lieber ML (1998) Confidence intervals for the receiver operating characteristic area in studies with small samples. Acad Radiol 5:561-571.

Randolph JJ, Edmondson RS (2005) Using the binomial effect size display (BESD) to present the magnitude of effect sizes to the evaluation audience. Practical Assessment, Research & Evaluation 10:1-7.

Rosenthal R, Rosnow RL (1985) Contrast analysis. Cambridge, UK: Oxford University Press.

Rosenthal R, Rosnow RL, Rubin DB (2000) Contrasts and effect sizes in behavioral research: a correlational approach. Cambridge: Cambridge University Press.

Rosenthal R, Rubin DB (1981) A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology 74:166-169.

Rosenthal R, Rubin DB (2003) r equivalent: a simple effect size indicator. Psychol Methods 8:492-496.

Smithson M (2003) Confidence intervals. Thousand Oaks: Sage Publications,Inc.

Stanislaw H, Todorov N (1999) Calculation of signal detection theory measures. Behav Res Methods Instrum Comput 31:137-149.

Steiger JH, Fouladi RT (1997) Noncentrality interval estimation and the evaluation of statistical models. In: What if there were no significance tests? (Harlow LL, Mulaik SA, Steiger JH, eds), pp 221-257. Mahwah, NJ: Erlbaum.

Stevens SS (1946) On the theory of scales of measurement. Science 103:677-680.