# G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences

**(in press). Behavior Research Methods.**

Franz Faul

Christian-Albrechts-Universität Kiel

Kiel, Germany

Edgar Erdfelder

Universität Mannheim

Mannheim, Germany

Albert-Georg Lang and Axel Buchner

Heinrich-Heine-Universität Düsseldorf

Düsseldorf, Germany

Please send correspondence to:

Prof. Dr. Edgar Erdfelder

Lehrstuhl für Psychologie III

Universität Mannheim

Schloss Ehrenhof Ost 255

D-68131 Mannheim, Germany

Email: erdfelder@psychologie.uni-mannheim.de

Phone +49 621 / 181 – 2146

Fax: + 49 621 / 181 - 3997

# Abstract

G*Power (Erdfelder, Faul, & Buchner, Behavior Research Methods, Instruments, & Computers, 1996) was designed as a general stand-alone power analysis program for statistical tests commonly used in social and behavioral research. G*Power 3 is a major extension of, and improvement over, the previous versions. It runs on widely used computer platforms (Windows XP, Windows Vista, Mac OS X 10.4) and covers many different statistical tests of the *t*-, *F*-, and $\chi^2$-test families. In addition, it includes power analyses for *z* tests and some exact tests. G*Power 3 provides improved effect size calculators and graphic options, it supports both a distribution-based and a design-based input mode, and it offers all types of power analyses users might be interested in. Like its predecessors, G*Power 3 is free.

# G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences

Statistics textbooks in the social, behavioral, and biomedical sciences typically stress the importance of power analyses. By definition, the power of a statistical test is the probability of rejecting its null hypothesis given that it is in fact false. Obviously, significance tests lacking statistical power are of limited use because they cannot reliably discriminate between the null hypothesis ($H_0$) and the alternative hypothesis ($H_1$) of interest. However, although power analyses are indispensable for rational statistical decisions, it took until the late 1980s until power charts (e.g., Scheffé, 1959) and power tables (e.g., Cohen, 1988) were supplemented by more efficient, precise, and easy-to-use power analysis programs for personal computers (Goldstein, 1989). G*Power 2 (Erdfelder, Faul, & Buchner, 1996) can be seen as a second-generation power analysis program designed as a stand-alone application to handle several types of statistical tests commonly used in social and behavioral research. In the past ten years, this program has been found useful not only in the social and behavioral sciences but also in many other disciplines that routinely apply statistical tests, for example, biology (Baeza & Stotz, 2003), genetics (Akkad et al., 2006), ecology (Sheppard, 1999), forest- and wildlife research (Mellina, Hinch, Donaldson, & Pearson, 2005), the geosciences (Busbey, 1999), pharmacology (Quednow et al., 2004), and medical research (Gleissner, Clusmann, Sassen, Elger, & Helmstaedter, 2006). G*Power 2 was evaluated positively in the reviews we are aware of (Kornbrot, 1997; Ortseifen, Bruckner, Burke, & Kieser, 1997; Thomas & Krebs, 1997), and it has been used in several power tutorials (e.g., Buchner, Erdfelder, & Faul, 1996, 1997; Erdfelder, Buchner, Faul & Brandt, 2004; Levin, 1997; Sheppard, 1999) as well as in statistics textbooks (e.g., Field, 2005; Keppel & Wickens, 2004; Myers & Well, 2003; Rasch, Friese, Hofmann, & Naumann, 2006a, 2006b). Nevertheless, the user feedback that we received converged with our own experience in showing some limitations and weaknesses of G*Power 2 that required a major extension and revision.

The present article describes G*Power 3, a program that was designed to address the problems of G*Power 2. We will first outline the major improvements in G*Power 3 (Section 1) before we discuss the types of power analyses covered by this program (Section 2). Next, we

describe program handling (Section 3) and the types of statistical tests to which it can be applied (Section 4). The fifth section is devoted to the statistical algorithms of G*Power 3 and their accuracy. Finally, program availability and some internet resources supporting users of G*Power 3 are described in Section 6.

## 1) Improvements in G*Power 3 compared to G*Power 2

G*Power 3 is an improvement over G*Power 2 in five major aspects. First, whereas G*Power 2 requires the DOS and Mac-OS 7-9 operating systems that were common in the 1990s but are now outdated, G*Power 3 runs on the currently most widely used personal computer platforms, that is, Windows XP, Windows Vista, and Mac OS X 10.4. The Windows and the Mac versions of the program are essentially equivalent. They use the same computational routines and share very similar user interfaces. For this reason, we will not differentiate between both versions in the sequel; users simply have to make sure to download the version appropriate for their operating system.

Second, whereas G*Power 2 is limited to three types of power analyses, G*Power 3 supports five different ways to assess statistical power. In addition to *a priori analyses*, *post hoc analyses*, and *compromise power analyses* that were already covered by G*Power 2, the new program also offers *sensitivity analyses* and *criterion analyses*.

Third, G*Power 3 provides dedicated power analysis options for a variety of frequently-used *t* tests, *F* tests, *z* tests, $\chi^2$ tests, and exact tests, rather than just the standard tests covered by G*Power 2. The tests captured by G*Power 3 are described in Section 3 along with their effect size parameters. Importantly, users are not limited to these tests because G*Power 3 also offers power analyses for generic *t*-, *F*-, *z*-, $\chi^2$, and binomial tests for which the noncentrality parameter of the distribution under $H_1$ may be entered directly. In this way, users are provided with a flexible tool for computing the power of basically any statistical test that uses *t*-, *F*-, *z*-, $\chi^2$, or binomial reference distributions.

Fourth, statistical tests can be specified in G*Power 3 using two different approaches, the distribution-based approach and the design-based approach. In the distribution-based approach,

users select (a) the family of the test statistic (i.e., *t*-, *F*-, *z*-, $\chi^2$, or exact test) and (b) the particular test within this family. This is the way in which power analyses were specified in G*Power 2. Additionally, a separate menu in G*Power 3 provides access to power analyses via the design-based approach: Users select (a) the parameter class the statistical test refers to (i.e., correlations, means, proportions, regression coefficients, variances) and (b) the design of the study (e.g., number of groups, independent vs. dependent samples, etc.). Based on the feedback we received to G*Power 2, we expect that some users might find the design-based input mode more intuitive and easier to use.

Fifth, G*Power 3 supports users with enhanced graphics features. The details of these features will be outlined in Section 3, along with a description of program handling.

## 2) Types of Statistical Power Analyses

The power (1-β) of a statistical test is the complement of β, which denotes the type-2 or beta error probability of falsely retaining an incorrect $H_0$. Statistical power depends on three classes of parameters: (1) the significance level (or, synonymously, the type-1 error probability) α of the test, (2) the size(s) of the sample(s) used for the test, and (3) an effect size parameter defining $H_1$ and thus indexing the degree of deviation from $H_0$ in the underlying population. Depending on the available resources, the actual phase of the research process, and the specific research question, five different types of power analysis can be reasonable (cf. Erdfelder et al., 2004; Erdfelder, Faul, & Buchner, 2005). We describe these methods and their uses in turn.

1) In *a priori power analyses* (Cohen, 1988), the sample size *N* is computed as a function of the required power level (1-β), the pre-specified significance level α, and the population effect size to be detected with probability (1-β). A priori analyses provide an efficient method of controlling statistical power *before* a study is actually conducted (e.g., Bredenkamp, 1969; Hager, 2006) and can be recommended whenever resources such as time and money required for data collection are not critical.

2) In contrast, *post hoc power analyses* (Cohen, 1988) often make sense after a study has already been conducted. In post hoc analyses, the power (1-β) is computed as a function of α, the

population effect size parameter, and the sample size(s) used in a study. It thus becomes possible to assess whether a published statistical test in fact had a fair chance to reject an incorrect $H_0$. Importantly, post-hoc analyses, like a priori analyses, require an $H_1$ effect size specification *for the underlying population*. They should not be confused with so-called retrospective power analyses in which the effect size is estimated from sample data and used to calculate the "observed power", a sample estimate of the true power[1]. Retrospective power analyses are based on the highly questionable assumption that the sample effect size is essentially identical to the effect size in the population from which it was drawn (Zumbo & Hubley, 1998). Obviously, this assumption is likely to be false, the more so the smaller the sample. In addition, sample effect sizes are typically biased estimates of their population counterparts (Richardson, 1996). For these reasons, we agree with other critics of retrospective power analyses (e.g., Gerard, Smith & Weerakkody, 1998; Hoenig & Heisey, 2001; Kromrey & Hogarty, 2000; Lenth, 2001; Steidl, Hayes, & Schauber, 1997). Rather than using retrospective power analyses, researchers should specify population effect sizes on a priori grounds. Effect size specification simply means to define the minimum degree of violation of $H_0$ a researcher would like to detect with a probability not less than $(1-\beta)$. Cohen's (1988) definitions of "small", "medium", and "large" effects can be helpful in such effect size specifications (see, e.g., Smith & Bayen, 2005). However, researchers should be aware of the fact that these conventions may have different meanings for different tests (cf. Erdfelder et al., 2005).

(3) In *compromise power analyses* (Erdfelder, 1984; Erdfelder et al., 1996; Müller, Manz, & Hoyer, 2002), both $\alpha$ and $1-\beta$ are computed as functions of the effect size, *N*, and an error probability ratio $q = \beta / \alpha$. To illustrate, $q = 1$ would mean that the researcher prefers balanced type-1 and type-2 error risks ($\alpha = \beta$), whereas $q = 4$ would imply that $\beta = 4 \cdot \alpha$ (cf. Cohen, 1988). Compromise power analyses can be useful both before and after data collection. For example, an a priori power analysis might result in a sample size that exceeds the available resources. In such a situation, a researcher could specify the maximum affordable sample size and, using a compromise power analysis, compute $\alpha$ and $(1-\beta)$ associated with, say, $q = \beta / \alpha = 4$. Alternatively, if a study has already been conducted but not yet been analyzed, a researcher could ask for a reasonable decision criterion that guarantees perfectly balanced error risks (i.e. $\alpha = \beta$), given the size of this

sample and a critical effect size she is interested in. Of course, compromise power analyses can easily result in unconventional significance levels larger than $\alpha = .05$ (in case of small samples or effect sizes) or less than $\alpha = .001$ (in case of large samples or effect sizes). However, we believe that the benefit of balanced type-1 and type-2 error risks often offsets the costs of violating significance level conventions (cf. Gigerenzer, Kraus, & Vitouch, 2004).

(4) In *sensitivity analyses* the critical population effect size is computed as a function of $\alpha$, $1-\beta$, and *N*. Sensitivity analyses may be particularly useful for evaluating published research. They provide answers to questions like "What is the effect size a study was able to detect with a power of $1-\beta = .80$, given its sample size and $\alpha$ as specified by the author? In other words, what is the minimum effect size the test was sufficiently sensitive to?" In addition, sensitivity analyses may be useful before conducting a study to see whether, given a limited *N*, the size of the effect that can be detected is at all realistic (or, for instance, way too large to be expected realistically).

(5) Finally, *criterion analyses* compute $\alpha$ (and the associated decision criterion) as a function of $1-\beta$, the effect size, and a given sample size. Criterion analyses are alternatives to post-hoc power analyses after a study has already been conducted. They may be reasonable whenever the control of $\alpha$ is less important than the control of $\beta$. In case of goodness-of-fit tests for statistical models, for example, it is most important to minimize the $\beta$-risk of wrong decisions in favor of the model ($H_0$). Researchers could thus use criterion analyses to compute the significance level $\alpha$ compatible with $\beta = .05$ for a small effect size.

Whereas G*Power 2 was limited to the first three types of power analysis, G*Power 3 now covers all five types. Based on the feedback we received from G*Power 2 users, we believe that any question related to statistical power that occurs in research practice can be translated into one of these analysis types.

## 3) Program Handling

Using G*Power 3 typically involves the following four steps: (1) Select the statistical test appropriate for your problem, (2) choose one of the five types of power analysis defined in the previous section, (3) provide the input parameters required for the analysis, and (4) click on "calculate" to obtain the results.

In the first step, the statistical test is chosen using the distribution-based or the design-based approach. G*Power 2 users probably have adapted to the distribution-based approach: One first selects the family of the test statistic (i.e., *t*-, *F*-, *z*-, $\chi^2$, or exact test) using the "Test family" menu in the main window. The "Statistical test" menu adapts accordingly, showing a list of all tests available for the test family. For the two groups *t* test, for example, one would first select the *t* family of distributions and then "Means: Differences between two independent means (two groups)" in the "Statistical test" menu (see Figure 1). Alternatively, one might use the design-based approach of test selection. With the "Tests" pull-down menu in the top row it is possible to select (a) the parameter class the statistical test refers to (i.e., correlations, means, proportions, regression coefficients, variances) and (b) the design of the study (e.g., number of groups, independent vs. dependent samples, etc.). For example, researchers would select "Means" ➔ "Two independent groups" to specify the two-groups *t* test (see Figure 2). The design-based approach has the advantage that test options referring to the same parameter class (e.g., means) are located in close proximity, whereas they may be scattered across different distribution families in the distribution-based approach.

---

Please insert Figures 1 and 2 about here.

---

In the second step the "Type of power analysis" menu in the center of the main window should be used to choose the appropriate analysis type. In the third step, the power analysis input parameters are specified in the lower left of the main window. To illustrate, an a priori power analysis for a two groups *t* test would require a decision between a one-tailed and a two-tailed test,

a specification of Cohen's (1988) effect size measure *d* under $H_1$, the significance level $\alpha$, the required power (1-$\beta$) of the test, and the preferred group size allocation ratio $n_2/n_1$. The final step consists of clicking "Calculate" to obtain the output in the lower right of the main window.

For instance, input parameters specifying a one-tailed *t* test, a medium effect size of $d = .5$, $\alpha = .05$, $(1-\beta) = .95$, and an allocation ratio of $n_2/n_1 = 1$ would result in a total sample size of $N=176$ (88 observation units in each group; see Figures 1 and 2). The noncentrality parameter $\delta$ defining the *t* distribution under $H_1$, the decision criterion to be used (i.e., the critical value of the *t* statistic), the degrees of freedom[2] of the *t* test and the actual power value are also displayed. Note that the actual power will often be slightly larger than the pre-specified power in a priori power analyses. The reason is that non-integer sample sizes are always rounded up by G*Power to obtain integer values consistent with a power level not less than the pre-specified one.

In addition to the numerical output, G*Power 3 displays the central ($H_0$) and the noncentral ($H_1$) test statistic distributions along with the decision criterion and the associated error probabilities in the upper part of the main window (see Figure 1)[3]. This supports understanding the effects of the input parameters and is likely to be a useful visualization tool in the teaching of, or the learning about, inferential statistics. The distributions plot may be printed, saved, or copied by clicking the right mouse button inside the plot area.

The input and output of each power calculation in a G*Power session is automatically written to a protocol that can be displayed by selecting the "Protocol of power analyses" tab in the main window. It is possible to clear the protocol, or to print, save, and copy the protocol in the same way as the distributions plot.

Because Cohen's (1988) book on power analysis appears to be well known in the social and behavioral sciences, we made use of his effect size measures whenever possible. Researchers not familiar with these measures and users preferring to compute Cohen's measures from more basic parameters can click on the "Determine" button to the left the effect size input field (see Figures 1 and 2). A drawer will open next to the main window and provide access to an effect size calculator tailored to the selected test (see Figure 2). For the two-groups *t* test, for example, users can specify the means ($\mu_1$, $\mu_2$) and the common standard deviation ($\sigma$) in the populations underlying the groups

to calculate Cohen's $d = |\mu_1-\mu_2|/\sigma$. Clicking the "Calculate and transfer to main window" button copies the computed effect size to the appropriate field in the main window.

---

Please insert Figure 3 about here.

---

Another useful option is the Power Plot window (see Figure 3) which is opened by clicking the "X-Y plot for a range of values" in the lower right corner of the main window (see Figures 1 and 2).

By selecting the appropriate parameters for the y- and the x-axis, one parameter ($\alpha$, power ($1-\beta$), effect size, or sample size) can be plotted as a function of any other parameter. Of the remaining two parameters, one can be chosen to draw a family of graphs, while the fourth parameter is kept constant. For instance, power ($1-\beta$) can be drawn as a function of the sample size for several different population effects sizes, keeping $\alpha$ at a particular value. The plot may be printed, saved, or copied by clicking the right mouse button inside the plot area. Selecting the "table" tab reveals the data underlying the plot; they may be copied to other applications.

The Power Plot window inherits all input parameters of the analysis that is active when the "X-Y plot for a range of values" button is pressed. Only some of these parameters can be directly manipulated in the Power Plot window. For instance, switching from a plot of a two-tailed test to that of a one-tailed test requires choosing the "Tail(s): One" option in the main window, followed by pressing the "X-Y plot for a range of values" button.

## 4) Types of Statistical Tests.

G*Power 3 provides power analyses for test statistics following *t*-, *F*-, $\chi^2$- or standard normal distributions under $H_0$ (either exact or asymptotic) and noncentral distributions of the same test families under $H_1$. In addition, it includes power analyses for some exact tests. In Tables 2 to 9 we briefly describe the tests currently covered by G*Power 3. Table 1 lists the symbols and their meanings as used in Tables 2 to 9.

Tests for Correlation and Regression

Table 2 summarizes the procedures supported for testing hypotheses on correlation and regression. One-sample tests are provided for the point-biserial model[4], i.e. correlations between a binary variable and a continuous variable, and for correlations between two normally distributed variables (Cohen, 1988, Chapter 3). The latter test uses the exact sample correlation coefficient distribution (Barabesi & Greco, 2002) or, optionally, a large sample approximation based on Fisher's *r*-to-*z* transformation. The two-sample test for differences between two correlations uses Cohen's effect size *q* (Cohen, 1988, Chapter 4) and is based on Fisher's *r*-to-*z* transformation. Cohen defines *q* = .10, *q* = .30, and *q* = .50 as "small", "medium", and "large" effects, respectively.

The two procedures available for the multiple regression model handle the cases of (a) a test of an overall effect, i.e. the hypothesis that the population value of $R^2$ is different from zero, and (b) a test of the hypothesis that adding additional predictors increases the value of $R^2$ (Cohen, 1988, Chapter 9). According to Cohen's criteria effects of size $f^2 = 0.02$, $f^2 = 0.15$, and $f^2 = 0.35$ are considered "small", "medium", and "large", respectively.

Tests for Means (univariate case)

Table 3 summarizes the power analysis procedures for tests on means. G*Power 3 supports all cases of the *t* test for means described by Cohen (1988, Chapter 2): the test for independent means, the test of the null hypothesis that the population mean equals some specified value (one sample case), and the test on the means of two dependent samples (matched pairs). Cohen's *d* and $d_z$ are used as effect size indices. Cohen defines *d* = 0.2, *d* = 0.5, and *d* = 0.8 as "small", "medium", and "large" effects, respectively. Effect size dialogs are available to compute the appropriate effect size parameter from means and standard deviations. For example, assume we want to compare visual search times for targets embedded in rare versus frequent local contexts in a within-subject design (cf. Hoffmann & Sebald, 2005, Exp. 1). It is expected that the mean search time for targets in rare contexts (say, 600 ms) should decrease by at least 10 ms in frequent contexts (i.e., to 590 ms) as a consequence of local contextual cuing. If prior evidence suggests population standard deviations

of, say, $\sigma = 25$ ms in each of the conditions and a correlation of $\rho = .70$ between search times in both conditions we can use the effect size drawer of G*Power 3 for the matched pairs *t* test to calculate the effect size $d_z = .516$ (see Table 3, row 2, for the formula). By selecting a post hoc power analysis for one-tailed matched pairs *t* tests, we easily see that for $d_z = .516$, $\alpha = .05$, and $N = 16$ participants the power is only $1-\beta = .47$. Thus, provided that the above assumptions are appropriate, the nonsignificant statistic $t(15) = 1.475$ obtained by Hoffmann and Sebald (2005, Exp. 1, p. 34) might in fact be due to a type 2 error. This interpretation would be consistent with the fact that Hoffmann and Sebald (2005) observed significant local contextual cuing effects in all other four experiments they reported.

The procedures provided by G*Power 3 to test effects in between-subjects designs with more than two groups (i.e., one-way ANOVA designs and general main effects and interactions in factorial ANOVA designs of any order) are identical to those in G*Power 2 (Erdfelder et al., 1996). In all these cases the effect size *f* as defined in Cohen (1988) is used. In a one-way ANOVA the effect size drawer can be used to compute *f* from the means and group sizes of *k* groups and a standard deviation common to all groups. For tests of effects in factorial designs, the effect size drawer offers the possibility to compute the effect size *f* from the variance explained by the tested effect and the error variance. Cohen defines $f = 0.1$, $f = 0.25$, and $f = 0.4$ as “small”, ”medium”, and ”large” effects, respectively.

New in G*Power 3 are procedures for analyzing main effects and interactions for $A \times B$ mixed designs, where *A* is a between-subjects factor (or an enumeration of the groups generated by cross-classification of several between-subject factors) and *B* is a within-subjects factor (or an enumeration of the repeated measures generated by cross-classification of several within-subject factors). Both the univariate and the multivariate approach to repeated measures (O’Brien & Kaiser, 1985) are supported. The multivariate approach will be discussed below. The univariate approach is based on the sphericity assumption. This assumption is correct if (in the population) all variances of the repeated measurements are equal and all correlations between pairs of repeated measurements are equal. If all the distributional assumptions are met, then the univariate approach is the most powerful method (Muller & Barton, 1989; O’Brien & Kaiser, 1985). Unfortunately, especially the assumption of equal correlations is often violated, which can lead to very misleading

results. In order to compensate for such adverse effects in tests of within effects or between-within interactions, the noncentrality parameter and the degrees of freedom of the *F*-distribution can be multiplied by a correction factor ε (Geisser & Greenhouse, 1958; Huynh & Feldt, 1970). ε is 1 if the sphericity assumption is met and approaches 1/(*m*-1) with increasing degrees of violation of sphericity, where *m* denotes the number of repeated measurements.

G*Power provides three separate yet very similar routines to calculate power in the univariate approach for between effects, within effects, and interactions. If the to-be-detected effect size *f* is known, these procedures are very easy to apply. To illustrate, Berti, Münzer, Schröger, and Pechmann (2006) compared the pitch discrimination ability of 10 musicians and 10 control subjects (between-subjects factor *A*) for 10 different interference conditions (within-subject factor *B*). Assuming that *A*, *B*, and $A \times B$ effects of "medium" size (*f* = .25, cf. Cohen, 1988; Table 3) should be detected given a correlation of ρ = .50 between repeated measures and a significance level of α = .05, the power values of the *F* tests for the *A* main effect, the *B* main effect, and the $A \times B$ interaction are easily computed as .30, .95, and .95, respectively, by inserting *f* = .25, α = .05, the total sample size (20), the number of groups (2), the number of repetitions (10) and ρ = .50 into the appropriate input fields of the procedures designed for these tests.

If the to-be-detected effect size *f* is unknown, it is necessary to compute *f* from more basic parameters characterizing the expected population scenario under $H_1$. To demonstrate the general procedure we will show how to do post-hoc power analyses in the scenario illustrated in Figure 4 assuming the variance and correlations structure defined in matrix $\mathbf{SR}_1$. We first consider the power of the within effect: We select the "*F* tests" family, the "ANOVA: Repeated measures, within factors" test, and "Post hoc" as the type of power analysis. Both the "Number of groups" and the "Repetitions" fields are set to 3. The "Total sample size" is set to 90 and the "α error probability" to 0.05. Referring to matrix $\mathbf{SR}_1$, we insert 0.3 in the input field "corr among rep measures", and -- since sphericity obviously holds in this case -- we set the "nonsphericity correction ε" to 1. To determine the effect size *f*, we first calculate $\sigma_\mu^2$, the variance of the within effect. From the three column means $\mu_{\bullet j}$ of matrix **M** and the grand mean $\mu_{\bullet\bullet}$ we get $\sigma_\mu^2$ = ((10-12.889)$^2$ + (13-12.889)$^2$ + (15.667-12.889)$^2$)/3 = 5.35679. Clicking the "Determine" button next to the effect size label opens the effect size drawer. We choose "From variances" and set the

"Variance explained by special effect" to 5.357 and the "Variance within groups" to $9^2 = 81$. Clicking the "Calculate and transfer to main window" button calculates an effect size $f = 0.2572$ and transfers $f$ to the effect size field in the main window. Clicking "Calculate" yields the results: The power is 0.997, the critical $F$ value with $df_1=2$ and $df_2=174$ is 3.048, and the noncentrality parameter $\lambda$ is 25.52. The procedure for tests of between-within interactions effects ("ANOVA: Repeated measures, within-between interaction") is almost identical to that just described. The only difference is in how the effect size $f$ is computed: Here we first calculate the variance of the residual values $\mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet}$ of matrix **M**: $\sigma_{\mu}^2 = ((10\text{-}10\text{-}15\text{+}12.889)^2 + \ldots + (12\text{-}15.667\text{-}11.333 + 12.889)^2)/9.0 = 1.90123$. Using the effect size drawer in the same way as above we get an effect size $f = 0.1532$ which results in a power of 0.653. To test between effects we choose "ANOVA: Repeated measures, between factors" and set all parameters to the same values as before. Note that in this case we do not need to specify $\varepsilon$—no correction is necessary because tests of between factors do not require the sphericity assumption. To calculate the effect size, we use "Effect size from means" in the effect size drawer. We select 3 groups, set "SD $\sigma$ within each group" to 9, and insert for each group the corresponding row mean $\mu_{i\bullet}$ of **M** (15,12.3333, 11.3333) and an equal group size of 30. An effect size $f = 0.1719571$ is calculated and the resulting power is 0.488.

Note that G*Power 3 can easily handle pure repeated measures designs without any between-subject factors (e.g., Frings & Wentura, 2005; Schwarz & Müller, 2006) by choosing the "ANOVA: Repeated Measures, within factors" procedure and setting the number of groups to 1.

Tests for Means (multivariate case)

G*Power 3 contains several procedures for performing power analyses in multivariate designs (see Table 3). All these tests belong to the $F$ test family.

The Hotelling $T^2$ tests are extensions of univariate $t$ tests to the multivariate case, where more than one dependent variable is measured: Instead of two single means two mean *vectors* are compared, and instead of a single variance, a variance-covariance matrix is considered (Rencher, 1998). In the one-sample case $H_0$ posits that the vector of population means is identical to a

specified constant mean vector. The effect size drawer can be used to calculate the effect size $\Delta$ from the difference $\vec{\mu} - \vec{c}$ and the expected variance-covariance matrix under $H_1$. For example, assume that we have two variables, a difference vector $\vec{\mu} - \vec{c}$ = {1.88, 1.88} under $H_1$, variances $\sigma_1^2 = 56.79$, $\sigma_2^2 = 29.28$, and a covariance of 11.98 (Rencher, 1998, p. 106). To perform a post hoc power analysis, choose "*F* tests", then "Hotellings $T^2$: One group mean vector" and set the analysis type to "Post hoc". Enter 2 in the "Response variables" field, and then press the "Determine" button next to effect size label. In the effect size drawer at "Input method: Means and…" choose "variance-covariance matrix" and press "Specify/Edit input values". Under the "Means" tab insert "1.88" in both input fields; under the "Cov Sigma" tab insert 56.79 and 29.28 in the main diagonal and 11.98 as off-diagonal element in the lower left cell. Pressing the button "Calculate and transfer to main window" initiates the calculation of the effect size (0.380) and transfers it to the main window. For this effect size, $\alpha = 0.05$, and a total sample size of $N = 100$ the power amounts to .9282. The procedure in the two-group case is exactly the same, with the following exceptions. First, in the effect size drawer two mean vectors have to be specified. Second, the group sizes may differ.

The MANOVA tests in G*Power 3 refer to the multivariate general linear model (O'Brien & Muller, 1993; O'Brien & Shieh, 1999): $\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon}$, where $\mathbf{Y}$ is $N \times p$ of rank $p$; $\mathbf{X}$ is $N \times r$ of rank $r$; and the $r \times p$ matrix $\mathbf{B}$ contains fixed coefficients. The rows of $\boldsymbol{\varepsilon}$ are taken to be independent $p$-variate normal random vectors with mean $\mathbf{0}$ and $p \times p$ positive-definite covariance matrix $\boldsymbol{\Sigma}$. The multivariate general linear hypothesis is $H_0$: $\mathbf{CBA} = \boldsymbol{\Theta}_0$, where $\mathbf{C}$ is $c \times r$ with full row rank, and $\mathbf{A}$ is $p \times a$ with full column rank (in G*Power 3, $\boldsymbol{\Theta}_0$ is assumed to be zero). $H_0$ has $df_1 = a \cdot c$ degrees of freedom. All tests of the hypothesis $H_0$ refer to the matrices

$$\mathbf{H} = N(\mathbf{CBU} - \boldsymbol{\Theta}_0)^T [\mathbf{C}(\ddot{\mathbf{X}}^T \mathbf{W} \ddot{\mathbf{X}})^{-1} \mathbf{C}^T]^{-1} (\mathbf{CBU} - \boldsymbol{\Theta}_0) = N\mathbf{H}^*$$

and

$$\mathbf{E} = \mathbf{U}^T \boldsymbol{\Sigma} \mathbf{U} (N - r_X),$$

where $\ddot{\mathbf{X}}$ is a $q \times q$ "essence model matrix", $\mathbf{W}$ is a $q \times q$ diagonal matrix containing weights $w_j = n_j / N$, and $\mathbf{X}^T\mathbf{X} = N(\ddot{\mathbf{X}}^T \mathbf{W} \ddot{\mathbf{X}})$ (see O'Brien & Shieh, 1999, p.14). Let $\{\phi_1^*, \ldots, \phi_s^*\}$ be the $s = \min(a, c)$ eigenvalues of $\mathbf{E}^{-1}\mathbf{H}^*$ and $\{\phi_1, \ldots, \phi_s\}$ the $s$ eigenvalues of $\mathbf{E}^{-1}\mathbf{H}/(N - r_X)$, i.e. $\phi_i = \phi_i^* N/(N - r_X)$.

G*Power 3 offers power analyses for the multivariate model following either the approach outlined in Muller and Peterson (1984; Muller, LaVange, Landesmann-Ramey & Ramey,1992) or alternatively the approach of O'Brien and Shieh (1999; Shieh, 2003). Both approaches approximate the exact distributions of Wilks *U* (Rao, 1951), the Hotelling-Lawley *T1* (Pillai & Samson, 1959), the Hotelling-Lawley *T2* (McKeon, 1974), and Pillai's *V* (Pillai & Mijares, 1959) by *F* distributions and are asymptotically equivalent. Table 5 outlines details of both approximations. The type of statistic (*U*, *T1*, *T2*, *V*) and the approach (Muller-Peterson or O'Brien-Shieh) can be selected in an Options dialog that can be evoked by clicking the button "Options" at the bottom of the main window.

The approach of Muller and Peterson (1984) has found widespread use; for instance, it has been adopted in the SPSS software package. We nevertheless recommend the approach of O'Brien and Shieh (1999) because it has a number of advantages: (a) Unlike the method of Muller and Peterson it provides the exact noncentral *F* distribution whenever the hypothesis involves at most $s = 1$ positive eigenvalues, (b) the approximations for $s > 1$ eigenvalues are almost always more accurate then the Muller and Peterson method (the Muller and Peterson method systematically underestimates power), and (c) it provides a simpler form of the noncentrality parameter, i.e. $\lambda = \lambda^* N$, where $\lambda^*$ is not a function of the total sample size.

G*Power 3 provides procedures to calculate the power for global effects in a "one-way MANOVA" and for special effects and interactions in factorial MANOVA designs. These procedures are the direct multivariate analogues of the ANOVA routines described above. Table 5 summarizes information that is needed in addition to the formulae given above to calculate the effect size *f* from hypothesized values for the mean matrix **M** (corresponding to matrix **B** in the model), the covariance matrix **Σ**, and the contrast matrix **C** describing the effect under scrutiny. The effect size drawer can be used to calculate *f* from known values of the statistic *U*, *T1*, *T2*, or *V*. Note, however, that the transformation of *T2* to *f* depends on the sample size. Thus this test statistic seems not very well suited for a priori analyses. In line with Bredenkamp and Erdfelder (1985) we recommend *V* as the multivariate test statistic.

Another group of procedures in G*Power 3 supports the multivariate approach to power analyses of repeated measures designs. G*Power provides separate but very similar routines for

the analysis of between effects, within effects and interactions in simple $A \times B$ designs, where *A* is a between-subjects factor and *B* a within-subjects factor. To illustrate the general procedure we describe in some detail a post hoc analysis of the within-effect for the scenario illustrated in Fig. 4 assuming the variance and correlations structure defined in matrix $\mathbf{SR}_2$. We first choose "*F* tests", then "MANOVA: Repeated measures, within factors". In the "Type of power analysis" menu we choose "Post hoc". We click the "Options" button to open a dialog in which we deselect the "Use mean correlation in effect size calculation" option. We choose the "Pillai V" statistic and the "O'Brien and Shieh" algorithm. Back to the main window we set both "Number of groups" and "Repetitions" to 3, the "Total sample size" to 90, and the "α error probability" to 0.05. To compute the effect size *f*(V) for the Pillai statistic we open the effect size drawer by clicking on the "Determine" button next to the effect size label. In the effect size drawer select, as procedure, "Effect size from mean and variance-covariance matrix" and, as input method, "SD and correlation matrix". Clicking on "Specify/Edit matrices" opens another window in which we specify the hypothesized parameters. In the "means" tab we insert our means matrix **M**, in the "Cov Sigma" tab we choose "SD and Correlation" and insert the values of $\mathbf{SR}_2$. Because this matrix is always symmetric, it suffices to specify the lower diagonal values. After closing the dialog and clicking on "Calculate and transfer to main window" we get a value 0.1791 for Pillai's V and the effect size *f*(V) = 0.4672. Clicking on "Calculate" shows that the power is 0.980. The analysis of between effects and interaction effects is done in an analogous way.

Tests for Proportions

The support for tests on proportions has been greatly enhanced in G*Power 3. Table 6 summarizes the tests that are currently implemented. In particular, all tests on proportions considered by Cohen (1988) are now available: (a) the sign test (Cohen, 1988, Chapter 5), (b) the *z*-test for the difference between two proportions (Cohen, 1988, Chapter 6), and (c) the $\chi^2$-test for goodness-of-fit and contingency tables (Cohen, 1988, Chapter 7).

The sign test is implemented as a special case ($c = 0.5$) of the more general binomial test (also available in G*Power 3) that a single proportion has a specified value *c*. In both procedures

Cohen's effect size *g* is used and exact power values based on the binomial distribution are calculated. Note, however, that due to the discrete nature of the binomial distribution the nominal value of $\alpha$ usually cannot be realized. Since the tables in Chapter 5 of Cohen's book use the $\alpha$ value closest to the nominal value, even if it is *higher* than the nominal value, the tabulated power values are sometimes larger than those calculated by G*Power 3. G*Power 3 always requires the actual $\alpha$ not to be larger than the nominal value.

Numerous procedures have been proposed to test the null hypothesis that two independent proportions are identical (D'Agostino, Chase & Belanger, 1988; Cohen, 1988; Suissa & Shuster, 1985; Upton, 1982), and G*Power 3 implements several of them. The simplest procedure is a *z* test with optional arcsin transformation and optional continuity correction. Besides these two computational options it can also be chosen whether Cohen's effect size measure *h* or, alternatively, two proportions are used to specify the alternate hypothesis. Using the options ("Use continuity correction" off, "Use arcsin transform" on) the procedure calculates power values close to those tabulated in Cohen (1988, Chapter 6). The options ("Use continuity correction" off, "Use arcsin transform" off) compute the uncorrected $\chi^2$-approximation (Fleiss, 1981) whereas ("Use continuity correction" on, "Use arcsin transform" off) computes the corrected $\chi^2$-approximation (Fleiss, 1981).

A second variant is Fisher's exact conditional test (Haseman, 1978). Normally, G*Power 3 calculates the exact unconditional power. However, despite the highly optimized algorithm used in G*Power 3, long computation times may result for large sample sizes (say $N > 1000$). Therefore a limiting *N* can be specified in the Options dialog that determines at which sample size G*Power 3 switches to a large sample approximation.

A third variant calculates the exact unconditional power for approximate test statistics *T* (Table 7 summarizes the supported statistics). The logic underlying this procedure is to enumerate all possible outcomes for the 2 x 2 binomial table, given fixed sample sizes $n_1$, $n_2$ in the two groups. This is done by choosing as success frequency $x_1$ and $x_2$ in each group any combination of the values $0 < x_1 \le n_1$ and $0 < x_2 \le n_2$. Given the success probabilities $\pi_1$, $\pi_2$ in each group, the probability of observing a table *X* with success frequencies $x_1$, $x_2$ is:

$$P(X \mid \pi_1, \pi_2) = \binom{n_1}{x_1} \pi_1^{x_1} (1-\pi_1)^{n_1 - x_1} \binom{n_2}{x_2} \pi_2^{x_2} (1-\pi_2)^{n_2 - x_2}$$

To calculate power (1-β) and the actual type-I error $\alpha^*$, the test statistic $T$ is computed for each table and compared with the critical value $T_\alpha$. If $A$ denotes the set of all tables $X$ rejected by this criterion, i.e. those with $T > T_\alpha$, then the power and the $\alpha$ level are given by: $1-\beta = \sum_{X \in A} P(X \mid \pi_1, \pi_2)$ and $\alpha^* = \sum_{X \in A} P(X \mid \pi_2, \pi_2)$, where $\pi_2$ denotes the success probability in both groups as assumed in the null hypothesis. Please note that the actual $\alpha$ level can be larger that the nominal level! The preferred input method (proportions, difference, risk ratio, odds ratio; see Table 6) and the test statistic to use (see Table 7) can be changed in the Options dialog. Note that the test statistic actually used to analyze the data should be chosen. For large sample sizes the exact computation may take too much time. Therefore, a limiting $N$ can be specified in the Options dialog that determines at which sample size G*Power switches to large sample approximations.

G*Power 3 also provides a group of procedures to test the hypothesis that the difference/risk ratio/odds ratio of a proportion with respect to a specified reference proportion $\pi$ is different under $H_1$ than a difference/risk ratio/odds ratio to the same reference proportion assumed in $H_0$. These procedures are available in the "exact" test family as "Proportions: Inequality (offset), two independent groups (unconditional)". The enumeration procedure described above for the tests on differences between proportions without offset is also used in this case. In the tests without offset, the different input parameters (differences, risk ratio, etc.) are equivalent ways of specifying two proportions. The specific choice has no influence on the results. In the case of tests with offset, however, each input method has a different set of available test statistics. The preferred input method (proportions, difference, risk ratio, odds ratio; see Table 6) and the test statistic to use (see Table 8) can be changed in the Options dialog. As in the other exact procedures, the computation may be time consuming and a limiting $N$ can be specified in the Options dialog that determines at which sample size G*Power switches to large sample approximations.

Also new in G*Power 3 is an exact procedure to calculate the power for the McNemar test. The null hypothesis of this test states that the proportions of successes are identical in two dependent samples. Figure 5 shows the structure of the underlying design: A binary response is sampled from the same subject or a matched pair in a standard and in a treatment condition. The

null hypothesis $\pi_s = \pi_t$ is formally equivalent to the hypothesis for the odd ratio: $OR = \pi_{12} / \pi_{21} = 1$. To fully specify $H_1$ we not only need to specify the odds ratio, but also the proportion $\pi_D$ of discordant pairs, that is, the expected proportion of responses that differ in the standard and the treatment condition. The exact procedure used in G*Power 3 calculates the unconditional power for the exact conditional test, which calculates the power conditional on the number $n_D$ of discordant pairs. Let $p(n_D = i)$ be the probability that the number of discordant pairs is $i$. Then the unconditional power is the sum over all $i \in \{0, \ldots. N\}$ of the conditional power for $n_D = i$ weighted with $p(n_D = i)$. This procedure is very efficient, but for very large sample sizes the exact computation nevertheless may take too much time. Again, a limiting $N$ can be specified in the Options dialog that determines at which sample size G*Power switches to a large sample approximation. The large sample approximation calculates the power based on an ordinary one sample binomial test with $Bin(N\pi_D, 0.5)$ as the distribution under $H_0$ and $Bin(N\pi_D, OR/(1+OR))$ as the $H_1$ distribution.

Tests for Variances

Table 9 summarizes important properties of the two procedures for testing hypotheses on variances that are currently supported by G*Power 3. In the one-group case the null hypothesis is tested that the population variance $\sigma^2$ has a specified value $c$. The variance ratio $\sigma^2/c$ is used as the effect size. The "central and noncentral" distributions, corresponding to $H_0$ and $H_1$, respectively, are *central* $\chi^2$ distributions with $N$–1 degrees of freedom (because $H_0$ and $H_1$ are based on the same mean). To compare the variance distributions under both hypotheses, the $H_1$ distribution is scaled with the value $r$ postulated for the ratio $\sigma^2/c$ in the alternate hypothesis, i.e. the noncentral distribution is $r\chi^2_{N-1}$ (Ostle & Malone, 1988). In the two-groups case $H_0$ states that the variances in two populations are identical ($\sigma_2/\sigma_1 = 1$). Analogous to the one sample case, two central $F$ distributions are compared, the $H_1$ distribution being scaled by the value of the variance ratio $\sigma_2/\sigma_1$ postulated in $H_1$.

Generic tests

Besides the specific routines described in Tables 2 to 9 that cover a considerable part of the tests commonly used, G*Power 3 provides "generic" power analysis routines that may be used for *any* test based on the *t*, *F*, $\chi^2$, *z*, and binomial distribution. In generic routines the parameters of the central and noncentral distributions are specified directly.

To demonstrate the uses and limitations of these generic routines we will show how to do a two-tailed power analysis for the one-sample *t* test by using the generic routine. You may compare the results with those of the specific routine available in G*Power for that test. First, we select the "t tests" family and then "Generic t test" (the generic test option is always located at the end of the list of tests). Next, we select "Post hoc" as the type of power analysis. We choose a two-tailed test and 0.05 as "α error probability". We now need to specify the noncentrality parameter δ and the degrees of freedom for our test. We look up the definitions for the one-sample test in Table 3 and find $\delta = d\sqrt{N}$ and $df = N - 1$. Assuming a "medium" effect of $d = 0.5$ and $N = 25$ we arrive at $\delta = 0.5 \cdot 5 = 2.5$, and $df = 24$. Inserting these values and clicking "Calculate" we obtain a power of $(1-\beta) = 0.6697$. The critical value $t = 2.0639$ corresponds to the specified α. In this post hoc power analysis, the generic routine is almost as simple as the specific routine. The main disadvantage of the generic routines is, however, that the dependence of the noncentrality parameter on the sample size is implicit. As a consequence, we cannot perform a priori analyses automatically. Rather, we need to iterate *N* by hand until we find an appropriate power value.

## 5) Statistical Methods and Numerical Algorithms

The subroutines used to compute the distribution functions (and the inverse) of the noncentral *t*, *F*, $\chi^2$, the *z* and the binomial distribution are based on the C-version of the DCDFLIB (available from http://www.netlib.org/random/) which was slightly modified for our purposes. G*Power 3 no longer provides approximate power analyses that were available in the speed mode of G*Power 2. Two arguments guided us in supporting exact power calculations only. First, four-digit precision of power calculations may be mandatory in many applications. For example, both compromise power analyses for very large samples and error probability adjustments in case of multiple tests of significance may result in very small values of α or β (Westermann & Hager, 1986). Second, as a

consequence of improved computer technology, exact calculations have become so fast that the speed gain associated with approximate power calculations is not even noticeable. Thus, from a computational standpoint, there is little advantage to using approximate rather than exact methods (cf. Bradley, Russel, & Reeve, 1998).

## 6) Program Availability and Internet Support

To summarize, G*Power 3 is a major extension of, and improvement over, G*Power 2 in that it offers easy-to-apply power analyses for a much larger variety of common statistical tests. Program handling is more flexible, easier to understand, and more intuitive compared to G*Power 2, reducing the risk of erroneous applications. The added graphical features should be useful for both research and teaching purposes. Thus, G*Power 3 is likely to become a useful tool for empirical researchers and students of applied statistics.

Like its predecessor, G*Power 3 is a noncommercial program that can be downloaded free of charge. Copies of the Mac and the Windows versions are available at http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/ only. Users interested in distributing the program in another way must ask for permission from the authors. Commercial distribution is strictly forbidden.

The G*Power 3 webpage will offer an expanding web-based tutorial describing how to use the program, along with examples. Users who let us know their e-mail addresses will be informed of updates. Although considerable effort has been put into program development and evaluation, there is no warranty whatsoever. Users are kindly asked to report possible bugs and difficulties in program handling to gpower-feedback@uni-duesseldorf.de.

# References

Akkad, D.A., Jagiello, P., Szyld, P., Goedde, R., Wieczorek, S., Gross, W.L., & Epplen, J.T. (2006). Promoter polymorphism rs3087456 in the MHC class II transactivator gene is not associated with susceptibility for selected autoimmune diseases in German patient groups. *International Journal of Immunogenetics*, *33*, 59-61.

Back, M.D., Schmukle, S.C., & Egloff, B. (2005). Measuring Task-Switching ability in the Implicit Association Test. *Experimental Psychology*, *52*, 167-179.

Baeza, J. A. & Stotz, W. (2003). Host-use and selection of differently colored sea anemones by the symbiotic crab Allopetrolisthes spinifrons. *Journal of Experimental Marine Biology and Ecology*, *284*, 25-39.

Barabesi, L. & Greco, L (2002). A note on the exact computation of the Student t Snedecor F and sample correlation coefficient distribution functions. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *51*, 105-110.

Berti, S., Münzer, S., Schröger, E., & Pechmann, T. (2006). Different interference effects in musicians and a control group. *Experimental Psychology*, *53*, 111-116

Bradley, D. R., Russell, R. L., & Reeve, C.P. (1998). The accuracy of four approximations to noncentral F. *Behavior Research Methods, Instruments, & Computers*, *30*, 478-500.

Bredenkamp, J. (1969). Über die Anwendung von Signifikanztests bei theorie-testenden Experimenten [The application of significance tests in theory-testing experiments]. *Psychologische Beiträge*, *11*, 275-285.

Bredenkamp, J., & Erdfelder, E. (1985). Multivariate Varianzanalyse nach dem V-Kriterium [Multivariate analysis of variance based on the V-criterion]. *Psychologische Beiträge*, *27*, 127-154.

Buchner, A., Erdfelder, E., & Faul, F. (1996). Teststärkeanalysen. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* [Handbook quantitative methods, pp. 123-136}. Weinheim, Germany: Psychologie Verlags Union.

Buchner, A., Erdfelder, E., & Faul, F. (1997). *How to Use G*Power* [WWW document]. URL http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/how_to_use_gpower.html

Busbey, A. B. I. (1999). Macintosh shareware/freeware earthscience software. *Computers & Geosciences*, *25*, 335-340.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ, Lawrence Erlbaum Associates.

D'Agostino, R.B, Chase, W. & Belanger, A. (1988). The appropriateness of some common procedures for testing the equality of two independent binomial populations. *The American Statistican*, 42,198-202.

Erdfelder, E. (1984). Zur Bedeutung und Kontrolle des beta-Fehlers bei der inferenzstatistischen Prüfung log-linearer Modelle [Significance and control of the beta error in statistical tests of log-linear models]. *Zeitschrift für Sozialpsychologie*, *15,* 18-32.

Erdfelder, E., Buchner, A., Faul, F., & Brandt, M. (2004). GPOWER: Teststärkeanalysen leicht gemacht. In E. Erdfelder & J. Funke (Eds.), *Allgemeine Psychologie und deduktivistische Methodologie* [Experimental psychology and deductive methodology, pp. 148-166]. Göttingen, Germany: Vandenhoeck & Ruprecht.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*, 1-11.

Erdfelder, E., Faul, F., & Buchner, A. (2005). Power analysis for categorical methods. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1565 – 1570). Chichester, GB: Wiley.

Farrington & Manning (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, 9, 1447-1454.

Field, A.P. (2005). Discovering statistics with SPSS (2nd ed.). London: Sage.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*, 2nd ed., New York: Wiley.

Frings, C. & Wentura, D. (2005). Negative priming with masked distractor-only prime trials: Awareness moderates negative priming. *Experimental Psychology*, *52*, 131-139.

Gerard, P.D., Smith, D.R., Weerakkody, G. (1998). Limits of retrospective power analysis. *Journal of Wildlife Management*, *62*, 801-807.

Gart, J.J. & Nam, J. (1988). Approximate interval estimation of the ratio in binomial parameters: A review and correction for skewness. Biometrics, 44, 323-338.

Gart, J.J. & Nam, J. (1990). Approximate interval estimation of the difference in binomial parameters: Correction for skewness and extension to multiple tables. Biometrics, 46, 637-643.

Geisser, S. & Greenhouse, S.W. (1958). An extension of Box's results on the use of the *F* distribution in multivariate analysis. *The Annals of Mathematical Statistics*, 29, 885-891.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 391-408). Thousand Oaks, CA: Sage.

Gleissner, U., Clusmann, H., Sassen, R., Elger, C.E., & Helmstaedter, C. (2006). Postsurgical outcome in pediatric patients with epilepsy: A comparison of patients with intellectual disabilities, subaverage intelligence, and average-range intelligence. *Epilepsia*, *47*, 406-414.

Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers. *The American Statistician*, *43*, 253-26.

Hager, W. (2006). Die Fallibität empirischer Daten und die Notwendigkeit der Kontrolle von falschen Entscheidungen [The fallibility of empirical data and the need for controlling for false decisions]. *Zeitschrift für Psychologie*, *214*, 10-23.

Haseman, J.K. (1978). Exact sample sizes for use with the Fisher-Irwin test for $2 \times 2$ tables. *Biometrics*, *34*, 106-109.

Hoenig, J.N. & Heisey, D.M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*, 19-24.

Hoffmann, J. & Sebald, A. (2005). Local contextual cuing in visual search. *Experimental Psychology*, *52*, 31-38.

Huynh, H. & Feldt, L.S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact *F*-distribution. *Journal of the American Statistical Association*, 65, 1582-1589.

Keppel, G. & Wickens, T.D. (2004). *Design and analysis. A researcher's handbook* (4th ed.). Upper Saddle River, NJ, Pearson Education International.

Kornbrot, D. E. (1997). Review of statistical shareware G*Power. *British Journal of Mathematical and Statistical Psychology*, *50*, 369-370.

Kromrey, J. & Hogarty, K.Y. (2000). Problems with probabilistic hindsight: A comparison of methods for retrospective statistical power analysis. *Multiple Linear Regression Viewpoints*, *26*, 7-14.

Lenth, R.V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*, 187-193.

Levin, J. R. (1997). Overcoming feelings of powerlessness in "aging" researches: A primer on statistical power in analysis of variance designs. *Psychology and Aging*, *12*, 84-106.

McKeon, J.J. (1974). *F* approximations to the distribution of Hotelling's $T_0^2$. *Biometrika*, 61, 381-383.

Mellina, E., Hinch, S.G., Donaldson, E.M., & Pearson, G. (2005). Stream habitat and rainbow trout (Oncorhynchus mykiss) physiological stress responses to streamside clear-cut logging in British Columbia. *Canadian Journal of Forest Research*, *35*, 541-556.

Miettinen, O. & Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine*, 4, 213-226.

Müller, J., Manz, R., & Hoyer, J. (2002). Was tun, wenn die Teststärke zu gering ist? Eine praktikable Strategie für Prä-Post-Designs. [What to do if the power is low? A useful strategy for pre-post designs]. *Psychotherapie, Psychosomatik, Medizinische Psychologie*, *52*, 408-416.

Muller, K.E. & Barton, C.N. (1989). Approximate power for repeated-measures ANOVA lacking sphericity. *Journal of the American Statistical Association*, 84, 549-555.

Muller, K.E., LaVange, L.M., Landesman-Ramey, S. & Ramey, C.T. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association*, 87,1209-1226.

Muller, K.E. & Peterson, B.L. (1984). Practical methods for computing power in testing the multivariate general linear hypothesis. *Computational Statistics and Data Analysis*, 2, 143-158.

Myers, J. L. & Well, A.D. (2003). *Research design and statistical analysis* (2nd ed.). Mahwah, NJ, Lawrence Erlbaum Associates.

O'Brien, R.G. & Kaiser, M.K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, 97, 316-333.

O'Brien, R.G., & Muller, K.E. (1993). Unified power analysis for t-tests through multivariate hypotheses. In Edwards, L.K. (Ed.). Applied analysis of variance in behavioral science. (pp. 297-344). New York, NY, US: Marcel Dekker.

O'Brien, R.G. & Shieh, G. (1999). Pragmatic, unifying algorithm gives power probabilities for common *F* tests of the multivariate general linear hypothesis. UnifyPow website: www.bio.ri.ccf.org/UnifyPow.

Ortseifen, C., Bruckner, T., Burke, M. & Kieser, M. (1997). An overview of software tools for sample size determination. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, 28, 91-118.

Ostle, B., & Malone, L.C. (1988). *Statistics in research: Basic Concepts and Techniques for Research Workers. Fourth Edition.* Ames, Iowa: Iowa State Press.

Pillai, K.C.S. & Mijares, T.A. (1959). On the moments of the trace of a matrix and approximations to its distribution. *Annals of Mathematical Statistics*, 30, 1135-1140.

Pillai, K.C.S. & Samson, P, Jr. (1959). On Hotelling's generalization of $T^2$. *Biometrika*, 46, 160-168.

Quednow, B.B., Kuhn, K.U., Stelzenmueller, R., Hoenig, K., Maier, W., & Wagner, M. (2004). Effects of serotonergic and noradrenergic antidepressants on auditory startle response in patients with major depression. *Psychopharmacology*, *175*, 399-406.

Rao, C.R. (1951). An asymptotic expansion of the distribution of Wilk's criterion. *Bulletin of the International Statistical Institute*, 33, 177-180.

Rasch, B., Friese, M., Hofmann, W.J., Naumann, E. (2006a). *Quantitative Methoden 1. Einführung in die Statistik* (2. Auflage) [Quantitative methods 1. Introduction to statistics]. Heidelberg: Springer

Rasch, B., Friese, M., Hofmann, W.J., Naumann, E. (2006b). *Quantitative Methoden 2. Einführung in die Statistik* (2. Auflage) [Quantitative methods 1. Introduction to statistics]. Heidelberg: Springer

Rencher, A.C. (1998). *Multivariate statistical inference and applications*. New York: John

Wiley & Sons.

Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods, Instruments, & Computers*, *28*, 12-22.

Scheffé, H. (1959). *The analysis of variance*. New York: John Wiley & Sons.

Schwarz, W. & Müller, D. (2006). Spatial associations in number-related tasks. A comparison of manual and pedal responses. *Experimental Psychology*, *53*, 4-15.

Sheppard, C. (1999). How large should my sample be? Some quick guides to sample size and the power of tests. *Marine Pollution Bulletin*, 38, 439-447.

Shieh, G. (2003). A comparative study of power and sample size calculations for multivariate general linear models. *Multivariate Behavioral Research*, 38, 285-307.

Smith, R.E. & Bayen, U.J. (2005). The effects of working memory resource availability on prospective memory: A formal modeling approach. *Experimental Psychology*, *52*, 243-256.

Steidl, R.J., Hayes, J.P., & Schauber, E. (1997). Statistical power analysis in wildlife research. *Journal of Wildlife Management*, *61*, 270-279.

Suissa, S. & Shuster, J.J. (1985). Exact unconditional sample sizes for $2 \times 2$ binomial trial. *Journal of the Royal Statistical Society, A*, 148, 317-327.

Thomas, L. & Krebs, C.J. (1997). A review of statistical power analysis software. *Bulletin of the Ecological Society of America*, 78, 126-139.

Upton, G.J.G. (1982). A comparison of alternative tests for the $2 \times 2$ comparative trial. *Journal of the Royal Statistical Society, A*, 145, 86-105.

Westermann, R. & Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics,*11, 117-146.

Zumbo, B.D. & Hubley, A.M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician*, *47*, 385-388.

# Footnotes

[1] The "observed power" is reported in many frequently used computer programs (e.g., the MANOVA procedure of SPSS).

[2] We recommend to check the df's reported by G*Power, for example, by comparing them to the df's reported by the program used to analyze the sample data. If the df's do not match, the input provided to G*Power is incorrect and the power calculations do not apply.

[3] Plots of the central and non-central distribution are only shown for tests based on the *t*-, *F*-, *z*-, $\chi^2$, or binomial distribution. No plots are shown for tests that involve an enumeration procedure (e.g. the McNemar test).

[4] We would like to thank Dave Kenny for making us aware of the fact the *t* test (correlation) power analyses of G*Power 2 are correct only in the point-biserial case (i.e., correlations between a binary variable and continuous variable, the latter being normally distributed for each value of the binary variable). For correlations between two continuous variables following a bivariate normal distribution, the *t* test (correlation) procedure of G*Power 2 overestimates power. For this reason, G*Power 3 offers separate power analyses for point-biserial correlations (in the *t* family of distributions) and correlations between two normally distributed variables (in the exact distribution family). However, power values will usually differ only slightly between procedures. To illustrate, assume we are interested in the power of a two-tailed test of $H_0$: $\rho = .00$ for continuously distributed measures derived from two Implicit Association Tests (IATs) differing in content. Assume further that, due to method-specific variance in both versions of the IAT, the true Pearson correlation is actually $\rho = .30$ (effect size). Given $\alpha = .05$ and $N = 57$ (see Back, Schmukle, & Egloff, 2005, Study 3, p. 173), an exact post-hoc power analysis for "Correlations: Differences from constant (one sample case)" reveals the correct power value of $(1-\beta) = .63$. Choosing the

incorrect “Correlation: point biserial model” procedure from the *t* test family would result in (1-β) = .65.

# Authors Note

Franz Faul, Institut für Psychologie, Christian-Albrechts-Universität, Kiel, Germany; Edgar Erdfelder, Lehrstuhl für Psychologie III, Universität Mannheim, Mannheim, Germany; Albert-Georg Lang and Axel Buchner, Institut für Experimentelle Psychologie, Heinrich-Heine-Universität, Düsseldorf, Germany.

Manuscript preparation has been supported by grants from the Deutsche Forschungsgemeinschaft (SFB 504, Project A12) and the state of Baden-Württemberg, Germany (Landesforschungsprogramm "Evidenzbasierte Stressprävention").

Correspondence concerning this article should be addressed to Franz Faul, Institut für Psychologie, Christian-Albrechts-Universität, Olshausenstr. 40, D-24098 Kiel, Germany (email: ffaul@psychologie.uni-kiel.de), or Edgar Erdfelder, Lehrstuhl für Psychologie III, Universität Mannheim, Schloss, D-68131 Mannheim, Germany (email: erdfelder@psychologie.uni-mannheim.de).

# Figure Captions

Figure 1:

The distribution-based approach of test specification in G*Power 3.0

Figure 2:

The design-based approach of test specification in G*Power 3.0 and the effect size drawer

Figure 3:

The Power Plot window of G*Power 3.0

Figure 4:

Sample 3 x 3 Repeated Measures designs: Three groups are repeatedly measured at three different times. The shaded portion of the table is the postulated matrix **M** of population means $\mu_{ij}$. The last column of the table contains the sample size in each group. The symmetric matrices $SR_i$ specify two different covariance structures between measurements taken at different times: The main diagonal contains the standard deviations of the measurements at each time, the off-diagonal elements contain the correlations between pairs of measurements taken at different times.

Figure 5:

Matched binary response design (McNemar test).

Figure 1

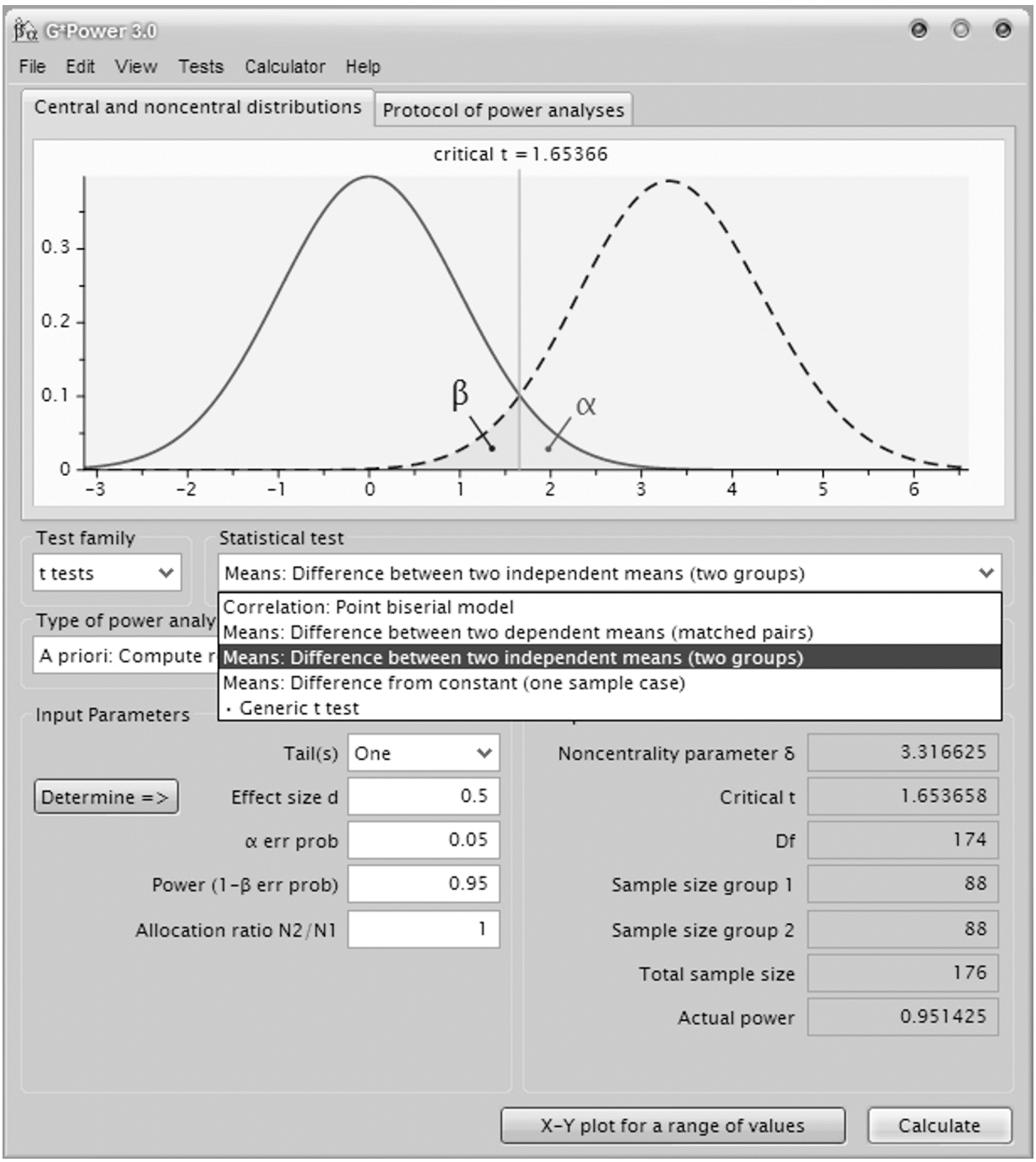
G*Power 3.0
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 1.65366
β
α
Test family
t tests
Statistical test
Means: Difference between two independent means (two groups)
Correlation: Point biserial model
Means: Difference between two dependent means (matched pairs)
Means: Difference between two independent means (two groups)
Means: Difference from constant (one sample case)
Generic t test
Type of power analy
A priori: Compute r
Input Parameters
Tail(s) One
Determine =>
Effect size d 0.5
α err prob 0.05
Power (1−β err prob) 0.95
Allocation ratio N2/N1 1
Noncentrality parameter δ 3.316625
Critical t 1.653658
Df 174
Sample size group 1 88
Sample size group 2 88
Total sample size 176
Actual power 0.951425
X-Y plot for a range of values
Calculate

Figure 2

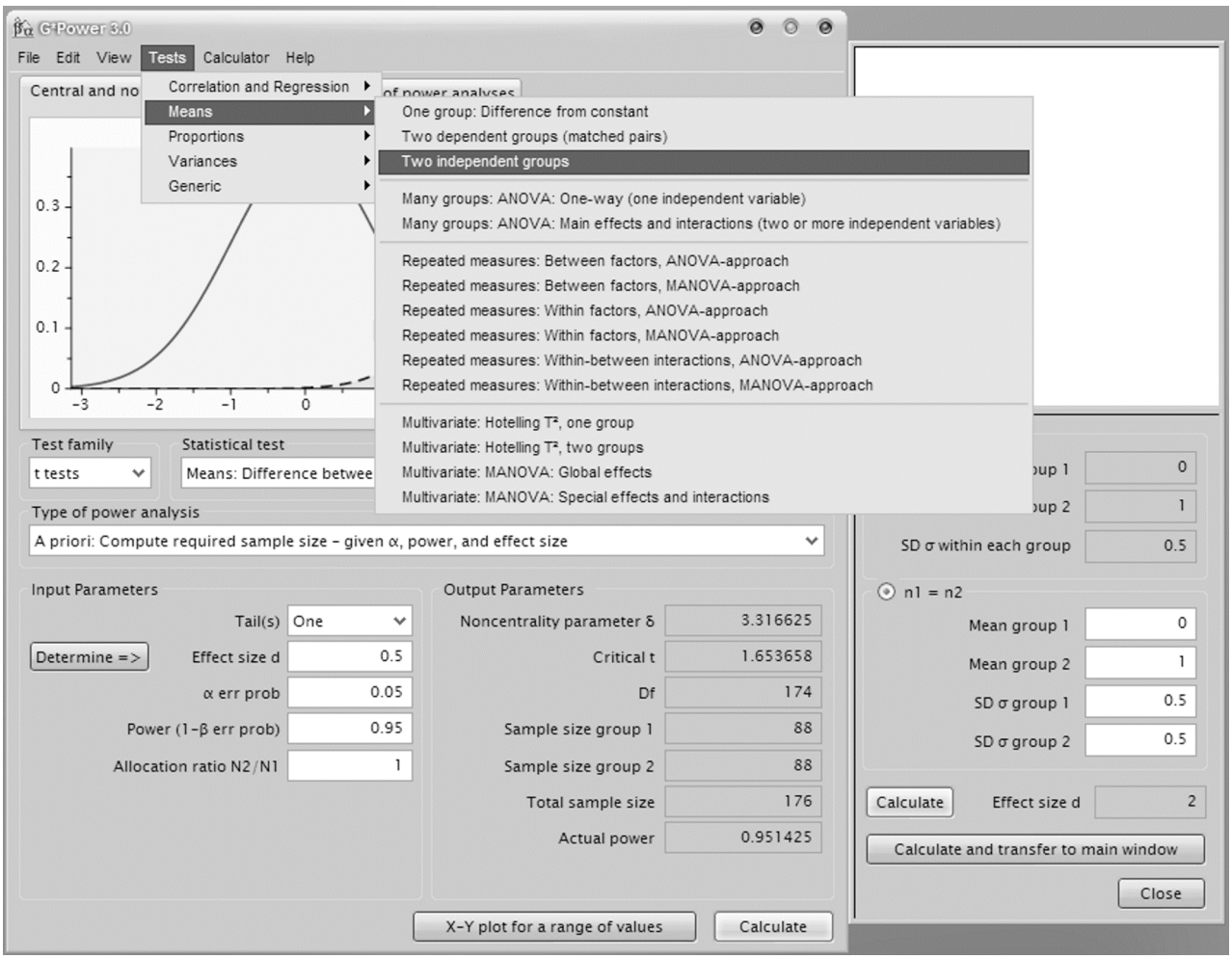
G*Power 3.0
File Edit View Tests Calculator Help
Correlation and Regression
Means
Proportions
Variances
Generic
One group: Difference from constant
Two dependent groups (matched pairs)
Two independent groups
Many groups: ANOVA: One-way (one independent variable)
Many groups: ANOVA: Main effects and interactions (two or more independent variables)
Repeated measures: Between factors, ANOVA-approach
Repeated measures: Between factors, MANOVA-approach
Repeated measures: Within factors, ANOVA-approach
Repeated measures: Within factors, MANOVA-approach
Repeated measures: Within-between interactions, ANOVA-approach
Repeated measures: Within-between interactions, MANOVA-approach
Multivariate: Hotelling T², one group
Multivariate: Hotelling T², two groups
Multivariate: MANOVA: Global effects
Multivariate: MANOVA: Special effects and interactions
Test family
t tests
Statistical test
Means: Difference betwee
Type of power analysis
A priori: Compute required sample size – given α, power, and effect size
Input Parameters
Tail(s) One
Determine =>
Effect size d 0.5
α err prob 0.05
Power (1–β err prob) 0.95
Allocation ratio N2/N1 1
Output Parameters
Noncentrality parameter δ 3.316625
Critical t 1.653658
Df 174
Sample size group 1 88
Sample size group 2 88
Total sample size 176
Actual power 0.951425
X-Y plot for a range of values
Calculate
SD σ within each group 0.5
n1 = n2
Mean group 1 0
Mean group 2 1
SD σ group 1 0.5
SD σ group 2 0.5
Calculate
Effect size d 2
Calculate and transfer to main window
Close

Figure 3

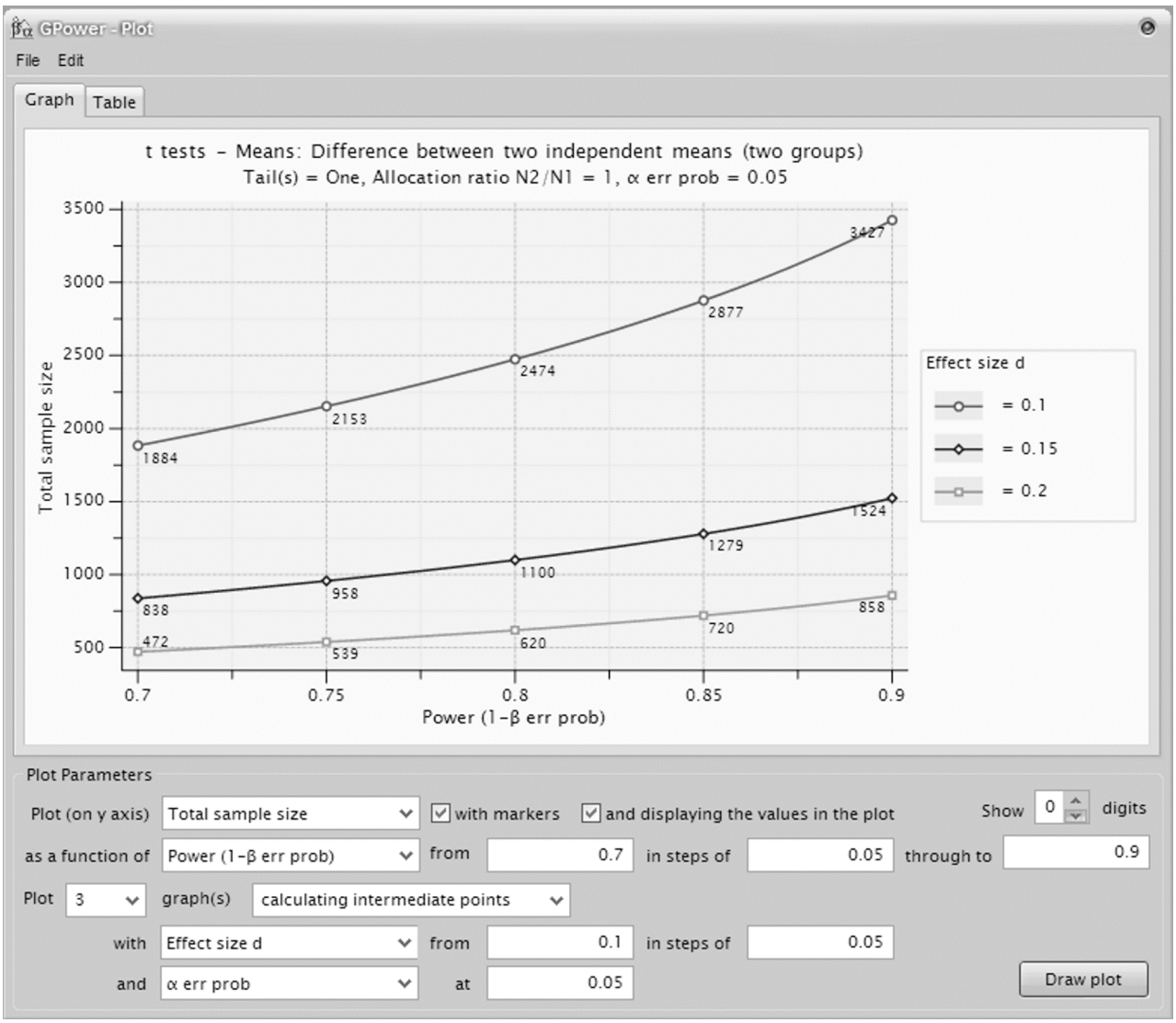
GPower - Plot
File Edit
Graph Table
t tests - Means: Difference between two independent means (two groups)
Tail(s) = One, Allocation ratio N2/N1 = 1, α err prob = 0.05
Total sample size
Power (1-β err prob)
Effect size d
= 0.1
= 0.15
= 0.2
Plot Parameters
Plot (on y axis) Total sample size with markers and displaying the values in the plot Show 0 digits
as a function of Power (1-β err prob) from 0.7 in steps of 0.05 through to 0.9
Plot 3 graph(s) calculating intermediate points
with Effect size d from 0.1 in steps of 0.05
and α err prob at 0.05
Draw plot

Figure 4

| | Time 1 | Time 2 | Time 3 | $\mu_{i\bullet}$ | $n_i$ |
|---|---|---|---|---|---|
| Group 1 | 10 | 15 | 20 | 15 | 30 |
| Group 2 | 10 | 12 | 15 | 12.333 | 30 |
| Group 3 | 10 | 12 | 12 | 11.333 | 30 |
| $\mu_{\bullet j}$ | 10 | 13 | 15.667 | $\mu_{\bullet\bullet}$ = 12.889 | |

$$\mathbf{SR}_2 = \begin{pmatrix} 10 & 0.3 & 0.1 \\ 0.3 & 9 & 0.3 \\ 0.1 & 0.3 & 8 \end{pmatrix} \quad \mathbf{SR}_1 = \begin{pmatrix} 9 & 0.3 & 0.3 \\ 0.3 & 9 & 0.3 \\ 0.3 & 0.3 & 9 \end{pmatrix}$$

Figure 5

| | Standard | | |
|---|---|---|---|
| Treatment | Yes | No | |
| Yes | $\pi_{11}$ | $\pi_{12}$ | $\pi_t$ |
| No | $\pi_{21}$ | $\pi_{22}$ | $1-\pi_t$ |
| | $\pi_s$ | $1-\pi_s$ | 1 |

Proportion of discordant pairs: $\pi_D = \pi_{12} + \pi_{21}$

Hypothesis: $\pi_s = \pi_t$ or equivalently $\pi_{12} = \pi_{21}$

Table 1: Symbols and their meanings as used in the tables in this article.

| Symbols | Meaning |
|---|---|
| $\mu, (\mu_i)$ | Population mean (in group *i*) |
| $\vec{\mu}, (\vec{\mu}_i)$ | Vector of population means (in group *i*) |
| $\mu_{x-y}$ | Population mean of the difference |
| $N$ | Total sample size |
| $n_i$ | Sample size in group *i* |
| $\sigma$ | Standard deviation in the population |
| $\sigma_\mu$ | Standard deviation of the effect |
| $\sigma_{x-y}$ | Standard deviation of the difference |
| $\lambda$ | Noncentrality parameter of the noncentral *F* and $\chi^2$ distribution |
| $\delta$ | Noncentrality parameter of the noncentral *t* distribution |
| $df$ | Degrees of freedom |
| $df_1, df_2$ | Numerator / denominator degrees of freedom |
| $\rho, (\rho_i)$ | Population correlation (in group *i*) |
| $R^2_{Y \cdot A}$, $R^2_{Y \cdot A,B}$ | Squared multiple correlation coefficients, corresponding to the proportion of *Y* variance that can be accounted for by multiple regression on the set of predictor variables *A* and $A \cup B$, respectively |
| $\mathbf{\Sigma}$ | Population variance-covariance matrix |
| $\mathbf{M}$ | Matrix of regression parameters (population means) |
| $\mathbf{C}$ | Contrast matrix (contrasts between rows of $\mathbf{M}$) |
| $\mathbf{A}$ | Contrast matrix (contrasts between columns of $\mathbf{M}$) |
| $\pi, (\pi_i)$ | Proportion (probability) of success (in group *i*) |

Table 2: Tests for correlation and regression

| Correlation and Regression | | | | | |
|---|---|---|---|---|---|
| Test | Test family | Null Hypothesis | Effect Size | Other Parameters | Noncentrality Parameter |
| Difference from zero: Point biserial model | *t* tests | $\rho = 0$ | $\rho$ | | $\delta = \sqrt{\frac{\rho^2}{1-\rho^2}} \cdot \sqrt{N}$<br>$df = N - 2$ |
| Difference from constant (bivariate normal) | exact | $\rho = c$ | $\rho$ | Constant correlation *c* | |
| Inequality of two correlation coefficients | *z* tests | $\rho_1 = \rho_2$ | $q = z_1 - z_2$<br>$z_i = \frac{1}{2} \ln \frac{1+\rho_i}{1-\rho_i}$ | | $m_1 = \frac{q}{s}$<br>$s = \sqrt{\frac{n_1 + n_2 - 6}{(n_1 - 3)(n_2 - 3)}}$ |
| Multiple Regression: deviation of $R^2$ from zero | *F* tests | $R^2_{Y \cdot A} = 0$ | $f^2 = \frac{R^2_{Y \cdot A}}{1 - R^2_{Y \cdot A}}$ | Number of predictors *p* (#A) | $\lambda = f^2 N$<br>$df_1 = p$<br>$df_2 = N - p - 1$ |
| Multiple Regression: increase of $R^2$ | *F* tests | $R^2_{Y \cdot A,B} = R^2_{Y \cdot A}$ | $f^2 = \frac{R^2_{Y \cdot A,B} - R^2_{Y \cdot A}}{1 - R^2_{Y \cdot A,B}}$ | Total number of predictors *p* (#A + #B)<br><br>Number of tested predictors *q* (#B) | $\lambda = f^2 N$<br>$df_1 = q$<br>$df_2 = N - p - 1$ |

Table 3: Tests for univariate means

| Means (univariate) | | | | | |
|---|---|---|---|---|---|
| Test | Test Family | Null Hypothesis | Effect Size | Other Parameters | Noncentrality Parameter and Degrees of Freedom |
| Difference from constant (one sample case) | *t* tests | $\mu = c$ | $d = \frac{\mu - c}{\sigma}$ | | $\delta = d\sqrt{N}$<br>$df = N - 1$ |
| Inequality of two dependent means (matched pairs) | *t* tests | $\mu_{x-y} = 0$ | $d_z = \frac{\mid \mu_{x-y} \mid}{\sigma_{x-.y}}$,<br>$\sigma_{x-y} = \sqrt{\sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y}$ | | $\delta = d_z\sqrt{N}$<br>$df = N - 1$ |
| Inequality of two independent means | *t* tests | $\mu_1 = \mu_2$ | $d = \frac{\mu_1 - \mu_2}{\sigma}$ | | $\delta = d\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$<br>$df = N - 2$ |
| ANOVA, fixed effects, one-way: Inequality of multiple means | *F* tests | $\mu_i - \mu = 0,$<br>$i = 1,\ldots,k$ | $f = \frac{\sigma_\mu}{\sigma},$<br>$\sigma_\mu^2 = \frac{\sum_{i=1}^{k} n_j(\mu_i - \mu)^2}{N}$ | Number of groups $k$ | $\lambda = f^2 N$<br>$df_1 = k - 1$<br>$df_2 = N - k$ |
| ANOVA, fixed effects, multi-factor designs and planned comparisons | *F* tests | $\mu_i - \mu = 0,$<br>$i = 1,\ldots,k$ | $f = \frac{\sigma_\mu}{\sigma}$ | Total number of cells in the design $k$<br><br>Degrees of freedom of the tested effect $q$. | $\lambda = f^2 N$<br>$df_1 = q$<br>$df_2 = (N - k)$ |
| ANOVA: repeated measurements, between effects | *F* tests | $\mu_i - \mu = 0,$<br>$i = 1,\ldots,k$ | $f = \frac{\sigma_\mu}{\sigma}$ | Levels of the between factor $k$<br><br>Levels of the repeated measure factor $m$<br><br>Population correlation among repeated measurements $\rho$<br><br>For within and within-between interactions: Nonsphericity correction $\varepsilon$<br>$\frac{1}{m-1} \le \varepsilon \le 1$ | $\lambda = f^2 u N \varepsilon$<br>$u = \frac{m}{1 + (m-1)\rho}$<br>$df_1 = (k - 1)$<br>$df_2 = (N - k)$ |
| ANOVA: repeated measurements within effects | *F* tests | $\mu_i - \mu = 0,$<br>$i = 1,\ldots,m$ | $f = \frac{\sigma_\mu}{\sigma}$ | | $\lambda = f^2 u N$<br>$u = \frac{m}{1 - \rho}$<br>$df_1 = (m-1)\varepsilon$<br>$df_2 = (N - k)(m-1)\varepsilon$ |
| ANOVA: repeated measurements between-within interactions | *F* tests | $\mu_{ij} - \mu_i - \cdots$<br>$\mu_j + \mu = 0,$<br>$i = 1,\ldots,k$<br>$j = 1,\ldots,m$ | $f = \frac{\sigma_\mu}{\sigma}$ | | $\lambda = f^2 u N \varepsilon$<br>$u = \frac{m}{1 - \rho}$<br>$df_1 = (k-1)(m-1)\varepsilon$<br>$df_2 = (N - k)(m-1)\varepsilon$ |

Table 4: Tests for multivariate means

<table>
<tr><th colspan="6">Means (multivariate)</th></tr>
<tr><th>Test</th><th>Test Family</th><th>Null Hypothesis</th><th>Effect Size</th><th>Other Parameters</th><th>Noncentrality Parameter and Degrees of Freedom</th></tr>
<tr><td>Hotelling $T^2$ difference from constant mean vector</td><td>F tests</td><td>$\vec{\mu} = \vec{c}$</td><td>$\Delta = \sqrt{\vec{v}^T \mathbf{\Sigma}^{-1} \vec{v}}$<br>$v = \vec{\mu} - \vec{c}$</td><td>Number of response variables k</td><td>$\lambda = \Delta^2 N$<br>$df_1 = k$<br>$df_2 = N - k$</td></tr>
<tr><td>Hotelling $T^2$ difference between two mean vectors</td><td>F tests</td><td>$\vec{\mu}_1 = \vec{\mu}_2$</td><td>$\Delta = \sqrt{\vec{v}^T \mathbf{\Sigma}^{-1} \vec{v}}$<br>$v = \vec{\mu}_1 - \vec{\mu}_2$</td><td>Number of response variables k</td><td>$\lambda = \Delta^2 \frac{n_1 n_2}{n_2 + n_2}$<br>$df_1 = k$<br>$df_2 = N - k - 1$</td></tr>
<tr><td>MANOVA global effects</td><td>F tests</td><td rowspan="2">CM = 0<br><br>Means matrix M<br><br>Contrast matrix C</td><td rowspan="4">Effect size<br><br>$f_{mult}$<br><br>depends on the test statistics:<br><br>▪ Wilks U<br>▪ Hotelling-Lawley T1<br>▪ Hotelling-Lawley T2<br>▪ Pillai V<br><br>and algorithms:<br><br>▪ Muller and Peterson (1984)<br>▪ O’Brien and Shieh (1999)</td><td>Number of groups g<br><br>Number of response variables k</td><td rowspan="2">Noncentrality parameter and degrees of freedom depend on the test statistic and algorithm used (see effect size column and Table 6).</td></tr>
<tr><td>MANOVA special effects</td><td>F tests</td><td>Number of groups g<br><br>Number of predictors p<br><br>Number of response variables k</td></tr>
<tr><td>MANOVA: repeated measurements, between effects</td><td>F tests</td><td rowspan="3">CMA = 0<br><br>Means matrix M<br><br>Between contrast matrix C<br><br>Within contrast matrix A</td><td rowspan="3">Levels of the between factor k<br><br>Levels of the repeated measure factor m</td><td rowspan="3"></td></tr>
<tr><td>MANOVA: repeated measurements within effects</td><td>F tests</td></tr>
<tr><td>MANOVA: repeated measurements between-within interactions</td><td>F tests</td></tr>
</table>

Table 5: Approximating univariate statistics for multivariate hypotheses

| Approximating univariate statistics | | | |
|---|---|---|---|
| Statistic | Formula | Numerator Degree of Freedom $df_2$ | Effect Size and Noncentrality Parameter |
| Wilks $U$ MP | $U = \prod_{k=1}^{s} (1+\phi_k)^{-1}$ | $df_2 = g(N-g_1)-g_2$ ; $g_1 = r+(a-c+1)/2$ $g_2 = (ca-2)/2$ $g = \begin{cases} 1 & ca \le 3 \\ \sqrt{\frac{(ca)^2-4}{c^2+a^2-5}} & ca \ge 4 \end{cases}$ | $f(U)^2 = \frac{1-U^{1/g}}{U^{1/g}}$ $\lambda = f(U)^2 df_2$ |
| Wilks $U$ OS | $U = \prod_{k=1}^{s} (1+\phi_k^*)^{-1}$ | | $f(U)^2 = \frac{1-U^{1/g}}{U^{1/g}}$ $\lambda = Ngf(U)^2$ |
| Pillai $V$ MP | $V = \prod_{k=1}^{s} \phi_k/(1+\phi_k)$ | $df_2 = s(N-r-a+s)$ | $f(V)^2 = \frac{V}{(s-V)}$ $\lambda = f(V)^2 df_2$ |
| Pillai $V$ OS | $V = \prod_{k=1}^{s} \phi_k^*/(1+\phi_k^*)$ | | $f(V)^2 = \frac{V}{(s-V)}$ $\lambda = Nsf(V)^2$ |
| Hotelling-Lawley *T1* MP | $T = \prod_{k=1}^{s} \phi_k$ | $df_2 = s(N-r-a-1)+2$ | $f(T)^2 = T/s$ $\lambda = f(T)^2 df_2$ |
| Hotelling-Lawley *T1* OS | $T = \prod_{k=1}^{s} \phi_k^*$ | | $f(T)^2 = T/s$ $\lambda = Nsf(T)^2$ |
| Hotelling-Lawley *T2* MP | $T = \prod_{k=1}^{s} \phi_k$ | $df_2 = 4+(ca+2)g$ $g = \frac{(N-r)^2-(N-r)g_4+g_3}{(N-r)g_2-g_1}$ $g_1 = c+2a+a^2-1$ $g_2 = c+a+1$ $g_3 = a(a+3)$ $g_4 = 2a+3$ $h = (df_2-2)/(N-r-a-1)$ | $f(T)^2 = T/h$ $\lambda = f(T)^2 df_2$ |
| Hotelling-Lawley *T2* OS | $T = \prod_{k=1}^{s} \phi_k^*$ | | $f(T)^2 = T/h$ $\lambda = Nhf(T)^2$ |

Table 6: Tests for proportions

| Proportions | | | | | |
|---|---|---|---|---|---|
| Test | Test Family | Hypothesis | Effect Size | Other Parameters | Non-centrality Parameter |
| Contingency tables & Goodness of Fit | $\chi^2$ tests | $\pi_{1i} = \pi_{0i}$<br>$i = 1,\ldots,k$<br>$\sum_{i=1}^{k} \pi_{0i} = 1$ | $w = \sqrt{\sum_{i=1}^{k} \frac{(\pi_{1i} - \pi_{0i})^2}{\pi_{0i}}}$ | | $\lambda = w^2 N$ |
| Difference from constant (one sample case) | exact | $\pi = c$ | $g = \pi - c$ | Constant proportion $c$ | |
| Inequality of two dependent proportions (McNemar) | exact | $\pi_{12} / \pi_{21} = 1$ | Odds ratio $= \pi_{12} / \pi_{21}$ | Proportion of discordant pairs $= \pi_{12} + \pi_{21}$ | |
| Sign test | exact | $\pi = 1/2$ | $g = \pi - 1/2$ | | |
| Inequality of two independent proportions | $z$ tests | $\pi_1 = \pi_2$ | (A) Alternate prop. $\pi_2$ ,<br>(B) $h = \phi_1 - \phi_2$<br>$\phi_i = 2\arcsin\sqrt{\pi_i}$ | (A) Null prop. $\pi_1$ | |
| Inequality of two independent proportions (Fisher's exact test) | exact | $\pi_1 = \pi_2$ | alternate prop. $\pi_1$ | Null prop. $\pi_2$ | |
| Inequality of two independent proportions (unconditional) | exact | $\pi_1 = \pi_2$ | (A) Alternate prop.: $\pi_1$<br>(B) Difference: $\pi_2 - \pi_1$<br>(C) Risk ratio: $\pi_2 / \pi_1$<br>(D) Odds ratio: $\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$ | Null prop. $\pi_2$ | |
| Inequality with offset of two independent proportions (unconditional) | exact | $\pi_1 = \pi_2 + c$ | (A) Alternate prop: $\pi_{1\|H_1}$<br>(B) Difference: $\pi_2 - \pi_{1\|H_1}$<br>(C) Risk ratio: $\pi_2 / \pi_{1\|H_1}$<br>(D) Odds ratio: $\frac{\pi_{1\|H_1}/(1-\pi_{1\|H_1})}{\pi_2/(1-\pi_2)}$ | (A) Prop.: $\pi_{1\|H_0}$<br>(B) Difference: $\pi_2 - \pi_{1\|H_0}$<br>(C) Risk ratio: $\pi_2 / \pi_{1\|H_0}$<br>(D) Odds ratio: $\frac{\pi_{1\|H_0}/(1-\pi_{1\|H_0})}{\pi_2/(1-\pi_2)}$<br>Null prop. $\pi_2$ | |

Table 7: Test statistics used in tests of the difference between two independent proportions. The *z*-tests in the table are more commonly known as $\chi^2$ tests (the equivalent *z* test is used to provide two-sided tests).

| Nr | Name | Statistic |
|---|---|---|
| 1 | *z*-test pooled variance | $z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\hat{\sigma}}\ ;\ \hat{\sigma} = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\ ;\ \hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1+n_2}$ |
| 2 | *z*-test pooled variance with continuity correction | $z = \frac{\hat{\pi}_1 - \hat{\pi}_2 + \frac{k}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}{\hat{\sigma}}$; $\hat{\sigma}$ see (1); $k = \begin{cases} -1 & \text{lower tail} \\ +1 & \text{upper tail} \end{cases}$ |
| 3 | *z*-test unpooled variance | $z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\hat{\sigma}}\ ;\ \hat{\sigma} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$ |
| 4 | *z*-test unpooled variance with continuity correction | $z = \frac{\hat{\pi}_1 - \hat{\pi}_2 + \frac{k}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}{\hat{\sigma}}$; $\hat{\sigma}$ see (3); $k = \begin{cases} -1 & \text{lower tail} \\ +1 & \text{upper tail} \end{cases}$ |
| 5 | Mantel-Haenszel test | $z = \frac{x_1 - E(x_1)}{\sqrt{V(x_1)}}\ ;\ E(x_1) = \frac{n_1(x_1+x_2)}{N}\ ; V(x_1) = \frac{n_1 n_2 (x_1+x_2)(N-x_1-x_2)}{N^2(N-1)}$ |
| 6 | Likelihood ratio (Upton, 1982) | $lr = 2\begin{pmatrix} t(x_1)+t(x_2)+t(1-x_1)+t(1-x_2)+t(N) \\ -t(n_1)-t(n_2)-t(x_1+x_2)-t(N-x_1-x_2) \end{pmatrix}$; $t(x) := x\ln(x)$ |
| 7 | *t* test with *df* = *N*-2 (D'Agostino et al, 1988) | $t_{N-2} = (x_1(1-x_2) - x_2(1-x_1))\sqrt{\frac{N-2}{N[n_2 x_1(1-x_1) + n_1 x_2(1-x_2)]}}$ |
| Note--- $x_i$ = success frequency in group *i*; $n_i$ = sample size in group *i*; $N = n_1 + n_2$ = total sample size; $\hat{\pi}_i = x_i / n_i$ | | |

Table 8: Test statistics used in tests of the difference with offset between two independent proportions.

| Nr | Name | Statistic |
|---|---|---|
| 1 | *z*-test pooled variance | $z = \frac{\hat{\pi}_1 - \hat{\pi}_2 - \delta}{\hat{\sigma}}$ ; $\hat{\sigma} = \sqrt{\hat{\pi}(1-\hat{\pi})(1/n_1 + 1/n_2)}$ ; $\hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2}$ |
| 2 | *z*-test pooled variance with cont.corr. | $z = \frac{\hat{\pi}_1 - \hat{\pi}_2 - \delta + k/2(1/n_1 + 1/n_2)}{\hat{\sigma}}$ ; $\hat{\sigma}$ see (1); $k = \begin{cases} -1 & \text{lower tail} \\ +1 & \text{upper tail} \end{cases}$ |
| 3 | *z*-test unpooled variance | $z = \frac{\hat{\pi}_1 - \hat{\pi}_2 - \delta}{\hat{\sigma}}$ ; $\hat{\sigma} = \sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_1 + \hat{\pi}_2(1-\hat{\pi}_2)/n_2}$ |
| 4 | *z*-test unpooled variance with cont.corr. | $z = \frac{\hat{\pi}_1 - \hat{\pi}_2 - \delta + k/2(1/n_1 + 1/n_2)}{\hat{\sigma}}$ ; $\hat{\sigma}$ see (3); $k = \begin{cases} -1 & \text{lower tail} \\ +1 & \text{upper tail} \end{cases}$ |
| 5 | *t* test with *df* = *N*-2 (D'Agostino et al, 1988) | $t_{N-2} = ((x_1 + \delta n_1)(1-x_2) - x_2(1-x_1-\delta n_1))K$; $K = \sqrt{(N-2)/\{N[n_2x_1(1-x_1) + n_1x_2(1-x_2)]\}}$ |
| 6 | Likelihood score ratio (difference)<br><br>(Miettinen & Nurminen, 1985)<br><br>Farrington & Manning (1990)<br><br>Gart & Nam (1990) | $z = \frac{\hat{\pi}_1 - \hat{\pi}_2 - \delta}{\hat{\sigma}}$ ; $\hat{\sigma} = \sqrt{(\tilde{\pi}_1(1-\tilde{\pi}_1)/n_1 + \tilde{\pi}_2(1-\tilde{\pi}_2)/n_2)K}$<br>Miettinen & Nurminen: *K* = N/(N-1), Farrington & Manning: *K* =1<br><br>$\tilde{\pi}_1 = 2u\cos(w) - b/(3a)$ ; $\tilde{\pi}_2 = \tilde{\pi}_1 - \delta$ ;<br>$\theta = n_2/n_1$ ; $a = 1+\theta$ ; $b = -(1+\theta+\hat{\pi}_1+\theta\hat{\pi}_2+\delta(\theta+2))$ ;<br>$c = \delta^2 + \delta(2\hat{\pi}_1 + \theta + 1) + \hat{\pi}_1 + \theta\hat{\pi}_2$ ; $d = -\hat{\pi}_1\delta(1+\delta)$;<br>$v = b^3/(3a)^3 - bc/(6a^2) + d/(2a)$ ; $w = (3.14159 + \cos^{-1}(v/u^3))/3$ ;<br>$u = \text{sgn}(v)\sqrt{b^2/(3a)^2 - c/(3a)}$<br><br>Skewness corrected *z'* (Gart & Nam); *z* according to Farrington & Manning:<br>$z' = \left(\sqrt{1 + 4\varphi(\varphi + z)} - 1\right)/2\varphi$ ; $V = (\tilde{\pi}_1(1-\tilde{\pi}_1)/n_1 + \tilde{\pi}_2(1-\tilde{\pi}_2)/n_2)^{-1}$ ;<br>$\varphi = V^{2/3}/6(\tilde{\pi}_1(1-\tilde{\pi}_1)(1-2\tilde{\pi}_1)/n_1 + \tilde{\pi}_2(1-\tilde{\pi}_2)(1-2\tilde{\pi}_2)/n_2)$ |
| 7 | Likelihood score ratio (risk ratio)<br><br>(Miettinen & Nurminen, 1985)<br><br>Farrington & Manning (1990)<br><br>Gart & Nam (1988) | $z = \frac{\hat{\pi}_1 - \hat{\pi}_2\phi}{\hat{\sigma}}$ ; $\hat{\sigma} = \sqrt{(\tilde{\pi}_1(1-\tilde{\pi}_1)/n_1 + \phi^2\tilde{\pi}_2(1-\tilde{\pi}_2)/n_2)K}$<br>Miettinen & Nurminen: *K* = N/(N-1), Farrington & Manning: K =1<br><br>$\tilde{\pi}_1 = \phi\tilde{\pi}_2$ ; $\tilde{\pi}_2 = \left(-b - \sqrt{b^2 - 4N\phi(x_1+x_2)}\right)/(2N\phi)$; $b = -(n_1+x_2)\phi - x_1 - n_2$ ;<br><br>Skewness corrected *z'* (Gart & Nam); *z* according to Farrington & Manning:<br>$z' = \left(\sqrt{1 + 4\varphi(\varphi + z)} - 1\right)/2\varphi$ ; $V = (1-\tilde{\pi}_1)/(\tilde{\pi}_1 n_1) + (1-\tilde{\pi}_2)/(\tilde{\pi}_2 n_2)$ ;<br>$\varphi = 1/(6V^{2/3})\left((1-\tilde{\pi}_1)(1-2\tilde{\pi}_1)/(n_1\tilde{\pi}_1)^2 + (1-\tilde{\pi}_2)(1-2\tilde{\pi}_2)/(n_2\tilde{\pi}_2)^2\right)$ |
| 8 | Likelihood score ratio (odds ratio)<br><br>(Miettinen & Nurminen, 1985) | $z = \frac{(\hat{\pi}_1 - \tilde{\pi}_1)/(\tilde{\pi}_1(1-\tilde{\pi}_1)) - (\hat{\pi}_2 - \tilde{\pi}_2)/(\tilde{\pi}_2(1-\tilde{\pi}_2))}{\sqrt{1/(n_1\tilde{\pi}_1(1-\tilde{\pi}_1)) + 1/(n_2\tilde{\pi}_2(1-\tilde{\pi}_2))K}}$<br><br>Miettinen & Nurminen: *K*=N/(N-1), Farrington & Manning: K =1<br><br>$\tilde{\pi}_1 = \tilde{\pi}_2\omega/(1 + \tilde{\pi}_2(\omega - 1))$ ; $\tilde{\pi}_2 = \left(-b + \sqrt{b^2 + 4a(x_1+x_2)}\right)/(2a)$;<br>$a = n_2(\omega - 1)$ ; $b = n_1\omega + n_2 - (x_1+x_2)(\omega - 1)$ |
| Note---$x_i$ = success frequency in group *i*; $n_i$ = sample size in group *i*; $N = n_1 + n_2$ = total sample size; $\hat{\pi}_i = x_i/n_i$ ; $\delta$ = difference between proportions postulated in $H_0$; $\phi$ = risk ratio postulated in $H_0$; $\omega$ = odds ratio postulated in $H_0$ | | |

Table 9: Tests for variances

| Variances | | | | | |
|---|---|---|---|---|---|
| Test | Test Family | Null Hypothesis | Effect Size | Other Parameters | Noncentrality Parameter |
| Difference from constant (one sample case) | $\chi^2$ tests | $\frac{\sigma^2}{c} = 1$ | Variance ratio<br>$r = \frac{\sigma^2}{c}$ | | $\lambda = 0$<br>($H_1$: central $\chi^2$ distribution, scaled with $r$)<br>$df = N - 1$ |
| Inequality of two variances | $F$ tests | $\frac{\sigma_2^2}{\sigma_1^2} = 1$ | Variance ratio<br>$r = \frac{\sigma_2^2}{\sigma_1^2}$ | | $\lambda = 0$<br>($H_1$: central $F$ distribution, scaled with $r$)<br>$df_1 = n_1 - 1$<br>$df_2 = n_2 - 1$ |