

## A Tour of UnifyPow: A SAS Module/Macro for Sample-Size Analysis

Ralph G. O'Brien, Cleveland Clinic Foundation

### ABSTRACT

UnifyPow is a freeware SAS® module/macro that performs statistical power analysis and other matters related to sample-size choice. Its functionality covers an extensive set of methods. One-group tests include the t and Wilcoxon for  $H_0: \mu = \mu_0$ ; the binomial and Z approximation for  $H_0: \pi = \pi_0$ ; t and Wilcoxon for paired means; McNemar's for paired proportions; and Fisher's r-to-Z for  $H_0: \rho = \rho_0$ . Two-group tests include t and Wilcoxon-Mann-Whitney for  $H_0: \mu_1 - \mu_2 = \delta_0$ ;  $\chi^2$ , likelihood ratio (LR), and Fisher's exact for  $H_0: \pi_1 = \pi_2$  (association in  $2 \times 2$  tables); r-to-Z for  $H_0: \rho_1 = \rho_2$ . J-group tests include ANOVA via the cell-means model with general linear contrasts;  $\chi^2$  and LR tests for proportions ( $2 \times J$  tables) with general linear contrasts on J logits. Also covered are the test of  $H_0: \beta_j = \beta_{j0}$  in a multiple regression model predicting Y from q X's ("Y | q X's") in which  $X_j$  has tolerance  $\text{Tol}(X_j) = 1 - R^2(X_j | \text{other } q - 1 \text{ X's})$ ; the comparison of  $R^2(Y | q_{\text{full}} \text{ X's})$  versus  $R^2(Y | q_{\text{reduced}} \text{ X's})$  from nested linear models; and  $G^2$  or  $-2\ln L$  from full vs. reduced logistic, log-linear, or Cox survival models. All methods handle unequal n designs. Any set of alpha-levels may be specified and results for directional (one-tailed) tests are given when potentially appropriate. Tabular output is well developed and graphical output is possible. A simple syntax unifies concepts and specifications across methods. For the current status of this ongoing project and information on downloading, see

<http://www.bio.ri.ccf.org/power.html>

A copy of this paper or its successor is available at this site as a PDF (Acrobat) file.

*Keywords:* power analysis, sample-size planning.

### INTRODUCTION

Collaborating statisticians should provide sound technical planning long before data are collected. Essential to this is choosing appropriate sample sizes and assessing statistical power. Nevertheless, sample-size planning has for too long been given short shrift by authors of texts on statistical methods, by statistics teachers, and by software developers, all of whom focus almost exclusively on methods related to analyzing data already collected. When attention *is* given to determining sample size and power, it is too often limited to crude approximation formulas or tables to handle a few elementary situations. As a result, a large proportion of research protocols still have inadequate or erroneous sample-size considerations. Nothing will address this problem better than the availability of good, affordable, comprehensive software, especially if this is functionally part of general statistical packages, such as the SAS System.

This need was reflected in the 1997 SASware Ballot.® 1246 votes were recorded for "provide power analysis and sample-size determination in all applicable procedures," ranking 8th out of the 58 choices in the SAS/STAT® section.

This report is an extension and update on my SUGI 22 presentation (O'Brien, 1997). UnifyPow is a major advancement over the SAS modules described in O'Brien and Muller (1993) and is now distributed via convenient Internet download. (See "Getting Stuff" at the end of this paper.) It comes as a single module that you just %INCLUDE in a normal SAS program. It can be converted to a true macro by just re-commenting about 10 lines of code, which are clearly designated. It is distributed as freeware, even though it rivals commercial applications costing hundreds of dollars. (Please pardon this shameless boast.)

Some main characteristics:

- *You do **not** need to be a SAS expert to use UnifyPow.* As you can judge yourself from the examples given herein, its input syntax is quite straightforward and its default output suffices in most cases.
- UnifyPow should run in any environment that supports the base SAS System. This includes MS Windows 3.1/95/NT, UNIX, Macintosh, CMS, MVS, OpenVMS, and OS/2. This gives UnifyPow outstanding portability.
- UnifyPow builds a SAS data set of its results. Thus, SAS users have option of developing customized reports, e.g., by writing their own PROC TABULATE or SAS/GRAPH® code, even merging results from two or more UnifyPow runs. This gives UnifyPow outstanding reporting flexibility.
- UnifyPow will handle unbalanced sample sizes in G independent groups for all relevant problems. It will compute both one- and two-tailed tests when appropriate. It will accept any alpha level. It will handle general contrasts (including  $df_H > 1$ ) on cell means, logits, and Fisher's Z-transformed correlations. This gives UnifyPow outstanding depth.
- UnifyPow avoids the use of antiquated "textbook" approximations if exact or virtually methods are feasible. This gives UnifyPow outstanding accuracy.

### WHAT CAN UnifyPow DO?

UnifyPow continues to evolve. This section describes what was released in December 1997, what I hope to release by August 1998, as well as some things that will appear later.

#### Key

- \* In December 1997 release.
- o Planned for August 1998 release.
- May not make August 1998 release.
- P Finds power for specified total sample size.

N Find total sample size for specified power.

CI(○) Sample-size analyses for confidence intervals: to assure with some probability, P, that the span of a  $(1 - \alpha)100\%$  CI will be less than some specified value. Both two-tailed and one-tailed intervals. *Will begin to appear in August 1998 release.*

**General Functionality**

- \* Runs in base SAS. Developed under UNIX and tested also under Windows 95, it should work “as is” in any SAS environment.
- \* All methods handle unbalanced sample sizes.
- \* Both two-tailed and one-tailed tests are considered whenever possible.
- \* Built-in tabling using PROC TABULATE macros.
- \* Results are collected in a SAS dataset, so experienced SAS users can customize the output. This includes using SAS/GRAPH.
- Options for automatic graphing using SAS/GRAPH.

**Specific Methods Supported**

**MEANS (LOCATION)**

- \* One-sample t test of  $H_0: \mu = \mu_0$ . Default:  $\mu_0 = 0$ . [P, N, CI(○)]
- \* One-sample Wilcoxon signed-rank test. [P, N]
- \* One-sample matched-pairs t test of  $H_0: \mu_1 - \mu_2 = \delta_0$ , plus corresponding Wilcoxon test. Default:  $\delta_0 = 0$ . [P, N, CI(○)]
- \* Two-sample standard t test of  $H_0: \mu_1 - \mu_2 = \delta_0$ . Default:  $\delta_0 = 0$ . [P, N, CI(○)]
- Two-sample Welch t test that allows unequal group variances. [P, N]
- \* Two-sample Wilcoxon-Mann-Whitney test. [P, N]
- \* Two-sample t test for matched-pairs data, i.e. testing  $H_0: (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) = \delta_0$ , plus the corresponding Wilcoxon test. Default:  $\delta_0 = 0$ . [P, N, CI(○)]
- \* One-way ANOVA overall F test on G independent means. [P, N]
- \* Cell-means model ANOVA: the general linear hypothesis on G independent means:  

$$H_0: C\mu = \mathbf{0},$$
 where C is  $q \times G$  and full row rank;  $q \geq 1$ . This handles virtually any common test on G cell means in a fixed effects design, including regular and special tests for factorial and nested designs. [P; N; CI, for  $q = 1$ ]
- \* G-sample tests for differences between matched-pairs data, including contrasts on these differences [P, N]

- Factorial ANOVA: tests of general main effects and interactions. Note: This can already be done by properly setting up  $H_0: C\mu = \mathbf{0}$ . [P, N]
- \* Situation in which t tests and ANOVAs are used for an outcome measure that is beta-binomial:  $Y_{ij}$  is binomial( $p_{ij}$ , N) where  $0 < p_{ij} < 1$  is a beta random variable with group mean  $\pi_j$  and standard deviation  $\sigma(p)$ . [P, N]
- \* General F test for linear models. First, outside of UnifyPow construct a set of *exemplary* data whose values conform to the expected values defined by some conjectured true model. That is, begin with an  $N_e \times q$  exemplary predictor (design) matrix,  $X_e$ , and a conjectured vector of regression coefficients,  $\beta$ . Then compute the exemplary outcome vector,  $y_e = X_e\beta$ . Next, use any ordinary linear models routine to fit and ‘test’ the exemplary dataset of  $N_e$  cases defined by  $y_e$  and  $X_e$ . [Note: the linear models routine must handle data with no residual variation (SSE = 0). PROC GLM does.] Let  $SSH_e$  be the obtained sums of squares hypothesis value. Using theory described in O’Brien and Muller (1993), UnifyPow will accept  $N_e$  and  $SSH_e$  values to drive a sample-size analysis. This technique is cumbersome, but it handles situations that fall outside of UnifyPow’s menu of common designs and tests for linear models. [P, N]
- General capability to handle repeated measures designs using the multivariate general linear model. Use of this particular feature may require users to have PROC IML installed on their system. Users not having IML will still be able to use all non-IML-based features, which is the vast majority of UnifyPow. [P, N]

**PROPORTIONS, LOGIT ANALYSIS, AND LOG-LINEAR MODELS**

- \* One-sample binomial test of  $H_0: \pi = \pi_0$ . Default:  $\pi_0 = 0.50$  (sign test). Calculations are virtually exact, relying on binomial calculations directly and not on the common Z approximation. Results include critical values in terms of number of “successes” out of N required for significance. [P, N, CI(○)].
- \* Unconditional test of two independent proportions,  $H_0: \pi_1 - \pi_2 = \delta_0$ . For default of  $\delta_0 = 0$ , this conforms to the usual  $\chi^2$  test and likelihood ratio test of association in a 2 (groups: A vs. B)  $\times$  2 (outcome: yes vs. no) contingency table. [P; N; CI(○) for both the log odds ratio,  $\ln(\psi) = \ln[\{\pi_1/(1 - \pi_1)\}/\{\pi_2/(1 - \pi_2)\}]$ , and  $\delta = \pi_1 - \pi_2$ .]
- \* Fisher's exact conditional test of two independent proportions (association in a 2  $\times$  2 table). [P, N]
- $\chi^2$  test and likelihood ratio test of overall association in a R  $\times$  C contingency table. [P, N]

\* Test of two correlated proportions. Let  $\pi_{ij}$  be the true proportion of cases falling into cell  $\{i, j\}$  of a  $2 \times 2$  contingency table, and  $\pi_{1+}$  and  $\pi_{+1}$  be the row and column marginal probabilities. Then  $H_0: \pi_{1+} - \pi_{+1} = 0$  is identical to  $H_0: \pi_{12} - \pi_{21} = 0$ . Conditioning on the discordant frequencies,  $f_{12}$  and  $f_{21}$ , gives McNemar's test. UnifyPow handles the more general form  $H_0: \pi_{12}/\pi_{21} = \psi_0$ , with default  $\theta_0 = 1.0$ . Powers are determined using virtually exact binomial calculations. [P, N, CI(○)]

\* Overall likelihood ratio test on G independent proportions or logits,  $\text{logit}(\pi_j) = \ln[\pi_j/(1 - \pi_j)]$ . [P, N]

\* Wald-type tests of the general linear hypothesis over G independent logits:

$$H_0: \mathbf{C}\boldsymbol{\psi} = \mathbf{0},$$

where  $\mathbf{C}$  is  $q \times G$  and full row rank,  $q \geq 1$ ;  $\boldsymbol{\psi}$  is the vector of the G logits. Used properly, this handles log-linear models testing applied to contingency tables in which one of the variables is a dichotomous “response” and the others are categorical predictors that form the G independent groups. [P; N; CI(○), for  $q = 1$ ]

\* General likelihood ratio test. Accepts  $-2\ln L(\text{full})$  and  $-2\ln L(\text{reduced})$ —or  $G^2(\text{full})$  and  $G^2(\text{reduced})$ —the log likelihood ratio statistics from two nested logistic regression or log-linear models (or other generalized linear model) that were fit to an “exemplary dataset” of  $N_e$  artificial cases constructed to give estimates identical to the user's conjectured population parameters. For log-linear models, see the theory of the strategy, the results of the Monte Carlo work, and the example in O'Brien (1986), which is reprised in Agresti (1990, pp 241-244). This strategy represents a pragmatic simplification of one later advanced by Self, Mauritsen, and Ohara (1992). Specifically, only their dominant term ( $\Delta$ ) is employed in UnifyPow's noncentrality calculation, because, as these authors concluded, the higher order term is “usually very close to zero.” ([P, N]

○ Overall likelihood ratio test for multiple logistic regression to predict a binary  $Y = 0$  or  $1$  from  $q$   $X$ 's. Let  $\pi = \text{Prob}[Y = 1] = E[Y]$ . An  $R^2$ -type measure (see Agresti, 1990, pp. 110) is

$$D(q \text{ X's}) = [2\ln L(q \text{ X's}) - 2\ln L(\text{null})]/[-2\ln L(\text{null})]$$

where  $-2\ln L(\text{null})$  and  $-2\ln L(q \text{ X's})$  are log-likelihood statistics from the null model (intercept only) and the one with  $q$  predictors (“ $q \text{ X's}$ ”). Let  $D'(q \text{ X's})$  be the population counterpart to  $D(q \text{ X's})$ , technically, the limit of  $D(q \text{ X's})$  as  $N_{\text{total}}$  increases. Power is computed by specifying  $\pi$  and  $D'(q \text{ X's})$ . [P, N]

### REGRESSION AND CORRELATION

\* One-sample test of the Pearson correlation,  $H_0: \rho = \rho_0$ . Default is  $\rho_0 = 0$  and the t test is used. When  $\rho_0 \neq 0$ , the test based on Fisher's r-to-Z transform is used. [P, N, CI(○)]

\* Overall F test for an ordinary least squares multiple regression model with  $q$   $X$ 's. Letting  $\rho^2$  be the population counterpart to the common  $R^2$  statistic, this tests  $H_0: \rho^2(q \text{ X's}) = 0$ . [P, N]

\* Usual t test of  $H_0: \beta_j = \beta_{0j}$  based on user's conjectures for  $\beta_j$ ;  $SD(X_j)$ , the standard deviation of  $X_j$ ;  $\text{Tol}(X_j) = 1 - R^2(X_j | \text{other } q - 1 \text{ X's})$ , the tolerance of  $X_j$  in the model with  $q$   $X$ 's; and  $SD(\epsilon)$ , the standard deviation of the residual variation term. [P, N, CI(○)]

\* Multiple partial correlation of  $q$  additional predictors given  $p$   $X$ 's already in an OLS regression model, i.e.,  $H_0: \rho^2(q \text{ X's} | p \text{ X's}) = 0$ . [P, N]

\* Linear contrast hypotheses on r-to-Z transformed independent Pearson correlations. Let the  $j^{\text{th}}$  element of  $\mathbf{z}$  be

$$Z(\rho_j) = 0.5 * \ln[(1 + \rho_j)/(1 - \rho_j)].$$

Then we can test

$$H_0: \mathbf{Cz} = \mathbf{0},$$

where  $\mathbf{C}$  is  $q \times G$  and full row rank;  $q \geq 1$ . By far, the most common use of this test is to compare two independent correlations,

$$H_0: \rho_1 - \rho_2 = 0,$$

using

$$H_0: Z(\rho_1) - Z(\rho_2) = 0.$$

[P, N]

– Capability to handle the univariate and multivariate general linear model. Use of this particular feature may require users to have PROC IML installed on their system. [P, N]

### SURVIVAL ANALYSIS

○ Log-rank tests comparing two groups. [P, N]

– Will offer greater functionality, but plans are not complete.

### AGREEMENT

– Test on  $\kappa$  statistic for  $2 \times 2$  table in one sample,  $H_0: \kappa = \kappa_0$ . There is some debate whether reasonable approximations have been developed for this. [P, N]

– Test on Lin's (correlation) coefficient of concordance between new assay and gold standard. Continuous measure; one group.

$H_0: K = K_0$ . See Lin (1992). [P, N]

**EXAMPLES**

These examples are intended to introduce UnifyPow’s easy-to-use syntax and show some of its depth. *They illustrate only a fraction of UnifyPow’s total functionality.* We start very simply in order to review basic concepts.

**Example 1: Testing One Proportion**

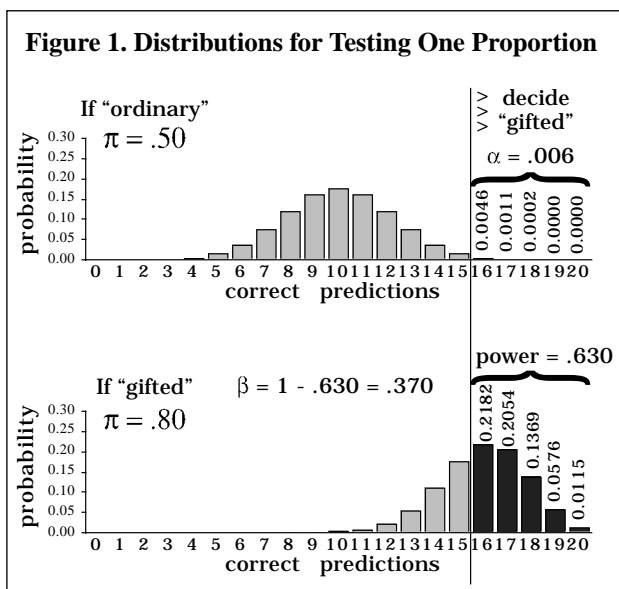
When “Mystic Michelle” was pregnant with her first child, she had eerie feelings that she could predict the future. So she tested herself in this by guessing the outcomes of fair coin tosses and now believes she is 80% accurate. She comes to Dr. Dee Bunker of the Center for Research in Applied Parapsychology to be certified as psychically “gifted,” so she can then seek fame and fortune.

Bunker designs the following experiment. Michelle will guess the outcomes of 20 coin tosses. If she gets 16 (80%) or more correct, then she will be certified. At first, this sounds quite reasonable to Michelle. But...

What are the Type I error rate ( $\alpha$ ) and power for this test? The probabilities for each possible outcome (0-20 correct) are graphed in Figure 1. The top distribution is simply the binomial distribution with  $N = 20$  and  $\pi = 0.50$ , the one that holds if Michelle is “ordinary.” The bottom distribution is binomial(20, 0.80) and corresponds to Michelle’s 80% claim. In one aspect, Bunker has set up a fairly stringent design: If Michelle is “ordinary” there is only a probability of  $\alpha = 0.006$  that she will correctly guess 16 or more correct and be erroneously certified. On the other hand, if Michelle really has the ability to correctly predict 80% of coin tosses in the long run, this  $N = 20$  experiment only has a probability of 0.630 of certifying her. This is the power of the test.

Let us consider how UnifyPow can be used to handle this problem. Consider the following statements:

```
pi .80                                     (Input 1a)
null .50 . (This is default.)
Ntotal 20 40
alpha .05 .01
```



All keywords are case insensitive. Note that each statement must begin on a new line and anything after an optional lone period ( . ) is a comment. A problem statement (here, PI) must come first. This one sets the design (one proportion) and the scenario ( $\pi = 0.80$ ). Following the problem statement, the others may come in any order. This NULL statement specifies  $H_0: \pi = 0.50$ , which is the default for this problem (ergo, giving the sign test) and thus could have been omitted. The NTOTAL and ALPHA statements specify that power calculations will be done for  $N = 20$  and  $N = 40$  crossed with  $\alpha = 0.05$  and  $\alpha = 0.01$ . Any  $\alpha$ -level may be used, thus allowing for easy analysis of Bonferroni tests, such as  $\alpha = .05/3 = 0.0167$  if a family of 3 tests was being co-protected. By default, results will be displayed for both one-tailed and two-tailed tests. The default output is given below.

**Output 1a. Power for Testing One Proportion**

Scenario: pi .80

		ALPHA			
		0.05		0.01	
		Total N		Total N	
		20	40	20	40
		Pow-er	Pow-er	Pow-er	Pow-er
Method	Type				
Exact Binomial	2-tail bnml	.804	.981	.630	.912
	1-tail bnml	.804	.992	.630	.957

**Output1b. Informaton Related to Testing One Proportion (edited to fit)**

Critical values and actual alpha levels using binomial distribution.

			ALPHA		
			0.01		
			Actual Alpha	Lower Crit Value	Upper Crit Value
Method	Total N	Type			
Exact Binomial	20	2-tail bnml	0.007	3	16
		1-tail bnml	0.006	.	16
	40	2-tail bnml	0.006	11	29
		1-tail bnml	0.008	.	28

Note that UnifyPow supplies the actual  $\alpha$  and the critical values for the number of successes. In this case, the one-tailed and two-tailed  $\alpha \leq 0.01$  tests have identical power (to 3 decimals), because each uses 16 or more for the upper critical region.

Students and clients understand such tables. They can see directly how increasing  $\alpha$ , increasing  $N$ , and using a one-tailed test can increase power. Mystic Michelle is actually quite sophisticated in statistics and decides to use UnifyPow herself to obtain the minimum sample sizes that will allow  $\alpha \leq 0.005$  and power  $\geq 0.99$  and power  $\geq 0.995$ :

```

/#                                     (Input 1b)
Same problem, but now find
minimum N to achieve specified
power at given alphas.
#/
pi .80
power .99 .995
alpha .005
tails 1
    
```

The key point here is the use of POWER rather than NTOTAL. This can be done for any kind of problem. Note as well how one can include comment blocks within UnifyPow statements and how the TAILS 1 statement causes only the results for the one-tailed tests to be tabled.

**Output 1c. Minimum N for Testing One Proportion**

Scenario: pi .80

		ALPHA	
		0.005	
		Minimum Power	
		.990	.995
		Total N	Total N
Method	Type		
Exact Binomial	1-tail bnml	61	69

**Output 1d. Information for Testing One Proportion**

Critical values and actual alpha levels using binomial distribution.

			ALPHA		
			0.005		
			Actual Alpha	Lower Crit Value	Upper Crit Value
Method	Minimum Power	Type			
Exact Binomial	.990	1-tail bnml	0.005	.	41
	.995	1-tail bnml	0.004	.	46

The output above tells us that if Michelle is ordinary ( $\pi = 0.50$ ) and if she were to make 69 predictions, there is only a 0.004 probability that she will be correct on 46 (67%) or more. If her general accuracy rate is really  $\pi = 0.80$ , she will achieve this result with probability at least 0.995.

UnifyPow uses exact or nearly exact calculations for the binomial and, therefore, its results may differ from those

given by other power analysis software, which use various approximations. For example, with  $\alpha = 0.05$  and  $N = 20$ , UnifyPow gives the exact power of 0.804. On the other hand, nQuery Advisor 2.0 ([www.statsolusa.com](http://www.statsolusa.com)) computes the power to be 0.903 and gives no warning message concerning the accuracy of its  $\chi^2$  approximation in this case. Power and Precision 1.20, also marketed as SamplePower by SPSS ([www.PowerAndPrecision.com](http://www.PowerAndPrecision.com)), gives 0.891, because it defaults to using an arcsin-based Normal approximation. Setting P&P's preference to "exact formula" does give the correct value, but this must be set each time you begin using this module. PASS 6.0 ([www.ncss.com/html/pass.html](http://www.ncss.com/html/pass.html)) uses the exact method, thus agreeing with UnifyPow.

**Example 2: Comparing Two Independent Proportions**

BeeBop Athletic Equipment is testing the "XDM-X," an experimental prototype of a running shoe being designed to be the successor to its popular XDM model. Does the XDM-X shoe reduce injuries? High-mileage runners will be randomly assigned to run either in the XDM-X or the XDM-S, which is just a standard XDM altered cosmetically to make it also look experimental. To get sufficient durability data on the XDM-X's, 2/3 of the runners will get the XDM-X. They will run at least 5 times/week, 50 miles/week for 26 weeks. One basic outcome measure will be: Did a serious running-related injury occur? Response: Yes/No.

Let  $\pi = \text{Pr}[\text{serious injury}]$ . Ample experience with the XDM shoe suggests that about 6% of these runners would experience such an injury in this time period. BeeBop believes that the XDM-X is a breakthrough in biomechanical engineering that could cut this rate in half, to 3%. Thus, the scenario is  $\pi_S = 0.06$  vs.  $\pi_X = 0.03$ . BeeBop plans to study about 200 runners, but could recruit as many as 270. What is the statistical power if  $\alpha = 0.05$  and  $\alpha = 0.01$ ?

The core UnifyPow statements are

```

pi .06 .03                                     (Input 2)
weight 1 2
NTotal 201 270
alpha .05 .01
    
```

This PI problem specifies a design with two independent groups. If the design was one testing three such proportions, then the problem statement would have been something like

```

pi .06 .03 .02
    
```

The WEIGHT statement specifies that  $1/(1 + 2) = 1/3$  of the cases will be in the first group. NTOTAL calls for power to be computed on 201 and 270 total cases (67+134 and 90+180). The ALPHA statement calls for 0.05 and 0.01 test sizes.

Running UnifyPow with these statements generates the results tabulated in Output 2. Although comparing two independent proportions seems like a simple problem, gifted statisticians have been debating the fine points for years and now almost 25 different methods have been

suggested just to get good p values. Obtaining power probabilities for these tests is an even tougher research problem. UnifyPow gives approximate powers corresponding to four tests. The first method given (“Approximate Unconditional  $\chi^2$ ”) corresponds to the ordinary Pearson  $\chi^2$  test, which can be approximated well by doing an ordinary t test (i.e., using a pooled variance) on  $Y = 0$  (no) or 1 (yes) data; see D’Agostino, Chase, and Belanger (1988). The second set of values is also based on the t, but uses the unpooled variance term and approximates the unconditional exact test (Suissa and Shuster, 1985; O’Brien and Muller, 1993). Purists should favor the Suissa-Shuster test, because it carries the “exact” moniker but does not rely on the questionable conditioning restriction of Fisher’s exact test, which is UnifyPow’s third method. Pragmatists will see from UnifyPow’s results that Fisher’s exact test generally has lower power than the others. Finally, the last approximation corresponds to testing this hypothesis via a standard likelihood-based logit analysis.

In this case, BeeBop agrees that these power values are much too small. After silently cursing the statistician who brought them this bad news, they consider an alternative outcome measure, considered next.

**Output 2. Power for Testing Two Independent Proportions.**

Scenario: pi .06 .03

		ALPHA			
		0.05		0.01	
		Total N		Total N	
		201	270	201	270
		Pow-er	Pow-er	Pow-er	Pow-er
Method	Type				
Approximate Unconditional "chi^2"	2-tld t apr	.175	.220	.060	.082
	1-tld t apr	.267	.323	.096	.126
Exact Unconditional**	2-tld t apr	.151	.187	.049	.065
	1-tld t apr	.234	.281	.079	.103
Fisher's exact conditional	2-tld aprx	.091	.129	.025	.040
	1-tld aprx	.153	.207	.044	.067
Likh Ratio for Log Odds Ratio	2-tail Z	.169	.211	.057	.077
	1-tail Z	.258	.311	.091	.120

\*The Approximate Unconditional corresponds to the Ordinary Pearson chi-square test for a 2 x 2 table. Technically, the method here uses a regular t test with  $Y = 0$  (no) or 1 (yes), which is known to offer more accurate p-levels and can be done with any standard t-test routine. See D’Agostino, Chase, and Belanger (1988), American Statistician, 1988, 42:198-202.

\*\*The Exact Unconditional corresponds to the test proposed by Suissa and Shuster (1985), J Royal Stat Soc A, 148:317-327).

**Example 3: Testing Two Means (Locations)**

In their study of the XDM-X running shoe, BeeBop decides to consider another outcome measure: the proportion of days a runner is injured, including days when he/she runs with the injury. Their data on the current XDM model suggests that this will have a median of about 9% and that 95% of these runners will have rates between 1% and 23%. BeeBop expects the XDM-X to improve on this, perhaps reducing the median injury-day rate to 7%, a 22% reduction. What is the statistical power for this outcome measure?

When the outcome measure itself is a proportion, it is common to transform it using an arcsin function before analysis by Normal-theory methods. We shall use  $Y_i = \arcsin(P_i^{1/2})$ , where  $P_i$  is the injury-day rate for the  $i^{th}$  runner. In the  $Y$  scale, the 9% median for  $P$  becomes a mean of  $\mu_Y = 0.30$  and the 95% limits on  $P$  (1% – 23%) become 0.10 and 0.50 for  $Y$ . Taking  $Y$  as Normal, this range covers about  $4\sigma$  units, so we conjecture that  $\sigma = 0.10$ . A median of 7% for  $P$  transforms to  $\mu_X = 0.27$ .

Consider the following UnifyPow statements:

```
mu .30 .27                                     (Input 3)
weight 1 2
SD .08 .10 .125
NTotal 201 270
Wilcoxon
methods all
```

Here we have a  $\mu$  problem, that is, we are comparing means with ANOVA methods. If we were comparing  $G$  means, then  $G$  values would have been given. The SD statement directs UnifyPow to examine the power over several possible standard deviations, assumed to be equal for the two groups. In this case, our target conjecture of  $\sigma = 0.10$  is being bracketed by  $\sigma = 0.08$  and  $\sigma = 0.125$ .

The WILCOXON statement directs UnifyPow to also give powers for a Wilcoxon-Mann-Whitney test, the standard nonparametric alternative to the two-group t test. Approximate powers are computed assuming that the parent distribution for  $Y$  is either Normal (“light” tailed:  $\gamma_2 = 0.0$ ), logistic (“slightly-heavy” tailed:  $\gamma_2 = 1.2$ ), or Laplace (“moderately-heavy” tailed:  $\gamma_2 = 3.0$ ). It is known that as kurtosis ( $\gamma_2$ ) increases, the WMW test becomes more powerful relative to the t test. UnifyPow can compute WMW power using three approximations, but only the default, Lehmann’s “3 moment” method (Lehmann, 1975), is shown here. This complex method, also given by Hettmansperger (1984), uses three nonparametric moments—namely the WMW effect size

$$p_1 = \text{Prob}(Y_{1i} - Y_{2i'} > \delta_0),$$

and two additional moments,  $p_2$ , and  $p_3$ , which are all described and tabled in UnifyPow’s output. Assessing location equality involves setting  $\delta_0 = 0$  and testing  $H_0: p_1 = 0.50$ . A side note: UnifyPow’s “Lehmann’s” approximation is not the rather crude and thus unsatisfactory “Lehmann’s” method studied by Lesaffre, Scheys, Fröhlich, and Bluhmki (1993).

The statement  
`method Noether`  
invokes Noether’s (1987) approximation, which is based only on  $p_1$ . `METHOD ARE` invokes the approximation based on the asymptotic relative efficiencies, which are used to transform the Normal-theory power for a t test to those of the WMW under the Normal, logistic, and Laplace parents, as well as a limiting case that gives a theoretical minimum of the power for the WMW. I made the Noether and ARE approximations optional after finding that Lehmann’s approximation generally did better in some Monte Carlo work I did to check all of this. `METHOD ALL` invokes all three approximations. To my knowledge, no other power analysis software uses the Lehmann-Hettesmansperger method or gives both the Noether and ARE methods, let alone all three. `nQuery Advisor` uses the Noether, but only for balanced designs and Normal parents. `Power and Precision` has no WMW functionality. `PASS` uses the ARE, but does not give the lower limits on power. In addition to the Normal, logistic, and Laplace parents, `PASS` also gives results for the uniform parent, a case I find too extreme for inclusion in `UnifyPow`.

The power results for this example are tabled in Output 3. Students and clients are often surprised to see how just a “minor” change in  $\sigma$  can substantially affect power.

There is also another way to specify the power scenario for the WMW test. While  $p_1$  can be found from  $\mu_1, \mu_2, \sigma$ , and the shape of the parent distribution, in some studies it is easier to just work with  $p_1$  directly. For example, it might be easiest to just ask Beebop researchers, “Consider pairing a random XDM-S runner and a random XDM-X runner. What is the probability that the XDM-S runner will have a greater proportion of injury days? If their conjecture is  $p_1 = 0.60$  and the parent distribution of Y is thought to be moderately-heavy tailed, then the following `UnifyPow` commands would be used:

```
2Wilcoxon .60
weight 1 2
parent Laplace
NTotal 201 270
```

Other values for `PARENT` are `NORMAL` and `LOGISTIC`.

**Related note.** The statement `2WILCOXON` has the counterpart `1WILCOXON` to handle the one-sample Wilcoxon test, otherwise known as the signed-rank test. Here, we are testing whether the median,  $\delta$ , exceeds some value,  $\delta_0$ . The effect size is

$$p_1 = \text{Prob}(Y > \delta_0).$$

### Output 3. Powers for Testing Two Means (Locations)

Scenario: mu .30 .27  
AND Alpha: 0.05

			Standard Deviation					
			0.08		0.1		0.125	
			Total N		Total N		Total N	
			201	270	201	270	201	270
			Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Method	Type	Parent						
Wilcoxon Mann-Whitney [Lehmann (p1, p2, p3) aprx]	2-tail W	Normal	.680	.806	.491	.616	.341	.437
		Logistic	.738	.856	.546	.675	.382	.488
		Laplace	.847	.935	.669	.796	.486	.610
	1-tail W	Normal	.785	.883	.618	.731	.464	.564
		Logistic	.832	.917	.669	.781	.508	.614
		Laplace	.912	.967	.776	.875	.612	.725
Wilcoxon Mann-Whitney [Noether (p1) aprx]	2-tail W	Normal	.675	.800	.492	.614	.343	.438
		Logistic	.732	.849	.545	.672	.384	.488
		Laplace	.839	.928	.666	.791	.487	.609
	1-tail W	Normal	.779	.876	.616	.727	.465	.563
		Logistic	.825	.911	.666	.776	.508	.613
		Laplace	.904	.962	.772	.870	.611	.723
Wilcoxon Mann-Whitney [aprx via ARE W vs. t]	2-tail W	Normal	.683	.807	.496	.619	.345	.440
		Logistic	.743	.858	.552	.679	.387	.492
		Laplace	.864	.944	.687	.810	.499	.622
		min ARE	.639	.767	.458	.576	.317	.406
	1-tail W	Normal	.787	.882	.621	.732	.467	.566
		Logistic	.834	.918	.673	.783	.512	.617
		Laplace	.922	.972	.790	.884	.624	.735
		min ARE	.750	.853	.584	.695	.437	.531
		min ARE	.750	.853	.584	.695	.437	.531
Ordinary t test	2-tail t	Normal	.703	.825	.514	.639	.358	.457
	1-tail t	Normal	.803	.895	.638	.750	.482	.583

Most applications of Wilcoxon’s signed rank test assess the difference between correlated means in a “matched pairs” situation. Here, Y is the difference score between pairs of related observations, so we usually set  $\delta_0 = 0$ . and test  $H_0: p_1 = 0.50$ . If you believed that 70% of the pairs would show a “positive” difference score and that this score would be a little more tail-heavy than the Normal, then you would specify

```
1Wilcoxon .70
parent logistic
```

Alternatively, one could define a `PairedMu` problem (not covered here) to set scenarios directly for the means, standard deviations, and within-pair correlation and then add “`WILCOXON`” to the `UnifyPow` input, as per Input 3.

**Example 4: One-Way ANOVA with Complex Contrasts**

What if BeeBop had 3 variations of their experimental shoe, XDM-X<sub>1</sub>, XDM-X<sub>2</sub>, XDM-X<sub>3</sub>? The design would have four groups. Expanding on Input 3, consider the statements:

```
mu .300 .275 .270 .265           (Input 4)
/#
2/3 of the runners will get some
version of the experimental shoe.
#/
weight 3 2 2 2
SD .08 .10 .125
NTotal 207 270
NoOverall
contrasts
"XDM-S vs. XDM-X (all versions)"
 3 -1 -1 -1
"Variation among XDM-X subtypes"
 0 1 -1 0
> 0 0 1 -1
```

If the NOOVERALL statement had not been used, UnifyPow would have computed the power for the overall F statistic with 3 degrees of freedom, a test of questionable value in this situation. The CONTRASTS statement focuses on the two main questions in this study: (1) Is the XDM-X (averaged over the three variations) better overall? (2) Is there any mean variation among the XDM-X subtypes? These are separate questions and do not need to be handled as a family of contrasts. Persons thinking more conservatively are free to use a Bonferroni adjustment, i.e.  $\alpha = 0.05/2 = 0.025$ , using

```
alpha .05 .025 . For 2 contrast family
Output 4 gives the results from Input 4.
```

**ANOVAs for Factorial Designs**

Being able to handle cell-means contrasts with multiple degrees of freedom gives UnifyPow exceptional flexibility, enough to handle virtually any fixed-effects ANOVA

situation, including factorial designs and nested designs, like in Example 4. For example, the generic 2 × 3 factorial design can be handled using the statements

```
mu #m11 #m12 #m13 #m21 #m22 #m23
weight #w11 #w12 #w13 #w21 #w22 #w23
SD # # . at least one
NTotal # # . at least one
NoOverall
contrasts
"A main" 1 1 1 -1 -1 -1
"B main" 1 -1 0 1 -1 0
> 0 1 -1 0 1 -1
"A by B" 1 -1 0 -1 1 0
> 0 1 -1 0 -1 1
```

#m12 and #w12 refer to the conjectured mean and the sample-size weight for cell {1, 2} of the design. The WEIGHT statement is always optional; balanced designs are default. In the future, users will be able to specify a factorial design that underlies the J means, so that the various standard main effects and interactions can be handled automatically. The syntax for the generic 2 × 3 might look something like this:

```
mu #m11 #m12 #m13 #m21 #m22 #m23
weight #w11 #w12 #w13 #w21 #w22 #w23
→ factorial "RowName" 2 "ColName" 3
SD # # . at least one value
NTotal # # . at least one value
```

A similar kind of functionality is planned for the PI problem, in order to handle common tests of association in contingency tables.

**Example 5: McNemar's Test of Correlated Proportions**

This example follows from a matched case-control study by Sartwell, et. al. (1969), which investigated whether taking oral contraceptives is associated with thromboembolism. Women of child-bearing age who were being treated with thromboembolism (the cases) where investigated to see whether they ever took The Pill

**Output 4. Powers for Testing 4 Means with Complex Contrasts**

Scenario: mu .300 .275 .270 .265

			Standard Deviation					
			0.08		0.1		0.125	
			Total N		Total N		Total N	
			207	270	207	270	207	270
			Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Test	Alpha	Type						
XDM-S vs. XDM-X (all versions)	0.05	2-tail t	.716	.825	.526	.639	.367	.457
		1-tail t	.813	.895	.649	.750	.491	.583
Variation among XDM-X subtypes	0.05	Regular F	.078	.087	.067	.073	.061	.065

(yes/no). Each thromboembolism case was matched to one female patient (the control) who had no history of thromboembolism, but who was similar on several other key variables.

Suppose in designing such a study we conjectured that our infinitely-large dataset would reveal the following type of association:

Thromboembolism Case Using The Pill?	Matched Control Using The Pill?	
	No	Yes
No	$\pi_{11} = 0.54$	$\pi_{12} = 0.08$
Yes	$\pi_{21} = 0.32$	$\pi_{22} = 0.06$

The null hypothesis is  $H_0: \pi_{12}/\pi_{21} = 1$  whereas the conjecture is that  $\pi_{12}/\pi_{21} = 1/4$ . This may be tested using McNemar's test. In essence, it ignores the diagonal cells and focuses only on the number of discordant pairs,  $N_D = f_{12} + f_{21}$ . Likewise, the effect size is only dependent on  $\pi_{12}$  and  $\pi_{21}$ . The UnifyPow statements for this problem are quite simple:

```
McNemar .32 .08 (Input 5)
alpha .01 .05
power .90 .95
```

Output 5 displays the results.

Obtaining the exact solution is computationally intensive, but machines are fast enough now to handle this with ease.  $N_D$  is a binomial random variable with  $N$  = total number of pairs and success rate  $\pi_D = \pi_{12} + \pi_{21}$ . Given  $N_D$ ,  $f_{12}$  is binomial with rate  $\pi' = \pi_{12}/(\pi_{12} + \pi_{21})$ . We test  $H_0: \pi' = 0.50$ . A little reflection shows that the exact unconditional power can be written as

$$\sum_{i=0}^N \{Pr[N_D=i] \cdot Power(N_D, \pi' | N_D=i)\}$$

UnifyPow computes this sum by beginning with the most probable values for  $N_D$  and stepping outward until virtually all of the probability mass on  $N_D$  is accounted for. Hence the calculation is virtually exact.

PASS 6.0 does not handle McNemar's test, but both nQuery Advisor and Power and Precision use approximations. nQA applies Miettinen's (1968) method and warns you if either  $f_{12}$  or  $f_{21}$  have expected values less than 3.8. P&P simply takes  $N_D$  to be a constant,  $N_D = N\pi_D$ , and then applies its ordinary binomial power calculations taking  $\pi'$  as the success rate. These approximations seem to perform satisfactorily. For a two-tailed test with  $\alpha = 0.05$ ,  $N = 200$ ,  $\pi_{12} = 0.08$ ,  $\pi_{21} = 0.02$  (so that  $E[f_{21}] = 4.0$ ), UnifyPow gives the exact value of 0.793, whereas nQA gives 0.803 and P&P (using the optional exact binomial computation) gives 0.804. Not bad, although one could certainly create extreme cases in which these approximations would break down.

**Other Examples**

Again, only a fraction of UnifyPow's capabilities are demonstrated here. Other examples can be found in the downloadable files.

**GETTING STUFF**

The UnifyPow source code, test files, and usage notes are distributed from the website

<http://www.bio.ri.ccf.org/power.html>

Ample instructions are given there. Please volunteer to register your name and email address, so I can send you announcements on new versions. Several documents, including this paper or its successor, are distributed as PDF files, which can be read and printed using Acrobat Reader, the marvelous free utility available for all common platforms at [www.adobe.com](http://www.adobe.com).

Output 5. Minimum Sample Sizes for McNemar's Test					
Scenario: McNemar .32 .08					
		ALPHA			
		0.05		0.01	
		Minimum Power		Minimum Power	
		.900	.950	.900	.950
		Pairs	Pairs	Pairs	Pairs
Method	Type				
McNemar (virtually exact)	2-tailed	67	82	94	111
	1-tailed	60	74	90	106

**Legal Disclaimer**

UnifyPow and its related files are freeware. You may give them to others at no charge, but you may not charge a “service fee” for their distribution. I distribute this work *pro bono*, in the spirit of collaborative science.

Use all of this at your own risk.

THIS FREeware COMES WITHOUT ANY WARRANTY WHATSOEVER. RALPH O'BRIEN AND THE CLEVELAND CLINIC FOUNDATION DO NOT AND CANNOT WARRANT THE PERFORMANCE OR RESULTS YOU MAY OBTAIN BY USING THIS FREeware AND ITS DOCUMENTATION. IN NO EVENT WILL RALPH O'BRIEN AND THE CLEVELAND CLINIC FOUNDATION BE LIABLE TO YOU FOR ANY CONSEQUENTIAL, INCIDENTAL, OR SPECIAL DAMAGES, INCLUDING LOST PROFITS OR LOST SAVINGS, EVEN IF RALPH O'BRIEN OR THE CLEVELAND CLINIC FOUNDATION HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES, OR FOR ANY CLAIM BY ANY THIRD PARTY.

**PROBLEMS AND SUGGESTIONS**

UnifyPow's computations are checked thoroughly using multiple methods: comparing its results to those obtained by using other software, to those in published tables and examples, and to results obtained from Monte Carlo simulation.

But no software this complex is ever free of problems and no software developer can check every situation that users might throw at it. I *really* do appreciate knowing if you encounter difficulties or have suggestions for improvements. Most additions to UnifyPow come about this way. On the other hand, I cannot promise to respond to matters that are mostly related to giving consulting advice. I correspond best via email.

**REFERENCES**

Agresti A (1990), *Categorical Data Analysis*, New York, John Wiley and Sons.

D'Agostino RB, Chase W, Belanger A (1988), “The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Proportions,” *American Statistician*, 42, 198-202.

Hettmansperger TP (1984), *Statistical Inference Based on Ranks*, New York: John Wiley and Sons.

Lehmann EL (1975), *Nonparametrics: Statistical Methods Based on Ranks*, New York: John Wiley and Sons.

Lesaffre E, Scheys I, Frölich J, Bluhmki E (1993), “Calculation of Power and Sample Size with Bounded Outcome Scores,” *Statistics in Medicine*, 12, 1063-1078.

Lin LI (1992), “Assay Validation Using the Concordance Correlation Coefficient,” *Biometrics*, 48, 599-604.

Noether GE (1987), “Sample Size Determination for Some Common NonParametric Tests,” *Journal of the American Statistical Association*, 82, 645-647.

Miettinen OS (1968), “On the Matched-Pairs Design in the Case of All-or-None Responses,” *Biometrics*, 24, 339-352.

O'Brien RG (1986), “Using the SAS System to Perform Power Analyses for Log-Linear Models,” *Proceedings of the Eleventh SAS Users Group International Conference, Cary, NC, SAS Institute*, 778-784.

O'Brien RG (1997), “UnifyPow: A SAS Macro for Sample-Size Analysis,” *Proceedings of the 22nd SAS Users Group International Conference, Cary, NC, SAS Institute*, 1353-1358.

O'Brien RG, Muller KE (1993), “Unified Power Analysis for t Tests through Multivariate Hypotheses,” in Edwards, L K (Ed.), *Applied Analysis of Variance in Behavioral Science*, New York: Marcel Dekker, 297-344.

Sartwell PE, Masi AT, Arthes FG, Greene GR, Smith HE (1969), “Thromboembolism and Oral Contraceptives: An Epidemiologic Case-Control Study,” *American Journal of Epidemiology*, 90: 365-380.

Self SG, Mauritsen RH, Ohara J (1992), “Power Calculations for Likelihood Ratio Tests in Generalized Linear Models,” *Biometrics*, 48, 31-39.

Suissa S, Shuster JJ (1985), “Exact Unconditional Sample Sizes for the  $2 \times 2$  Comparative Trial,” *Journal of the Royal Statistical Society (A)*, 148, 317-327.

---

SAS, SAS/GRAPH, SAS/STAT, and SASware Ballot are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.

CMS, MVS, and OS/2 are registered trademarks or trademarks of International Business Machines Corporation. ® indicates USA registration.

Other brand names and product names are registered trademarks or trademarks of their respective companies.

*Ralph O'Brien, PhD, loves to think that he directs the Collaborative Biostatistics Center within the Department of Biostatistics and Epidemiology, Cleveland Clinic Foundation, Cleveland, Ohio, 44195.. But in reality, this group of 25 terrific professionals directs him..*

216-445-9451; [robrien@bio.ri.ccf.org](mailto:robrien@bio.ri.ccf.org)  
<http://www.bio.ri.ccf.org/Resume/Pages/robrien.html>