(Some of the) UnifyPow Workshop Examples

Ralph G. O'Brien, Cleveland Clinic Foundation

Ordinary 2-group t-test comparing two independent means

Dr. Seth Alalgia[1] is planning a study assess whether a new form of biofeedback therapy can reduce, at least in the short term, the frequency and severity of chronic vascular headaches (cf. Blanchard et al., 1990). Specifically, he plans to conduct a double-blind, randomized trial in which patients will receive either an enhanced thermal (ET) biofeedback therapy or a sham placebo (SP) that simply gives non-contingent (essentially random) feedback to the patient. Each patient will be studied from a Monday to a Friday. On Monday morning, patients will be admitted to the university's General Clinical Research Center (GCRC) to begin identical, standard medical therapy consisting of 5μg/kg/day of litosamine. On Monday and Tuesday evenings, patients will complete an extensive questionnaire that yields the pretreatment score on the Vascular Headache Index (PreVHI). On Wednesday morning, patients will be randomized to ET or SP groups, using the minimization scheme of Pocock and Simon (1975) to minimize group differences on PreVHI, gender, and age strata. The main behavioral therapy session will be done on the Wednesday afternoon, with a followup session early Thursday morning. Questionnaire data from Thursday and Friday will yield the PostVHI measure.

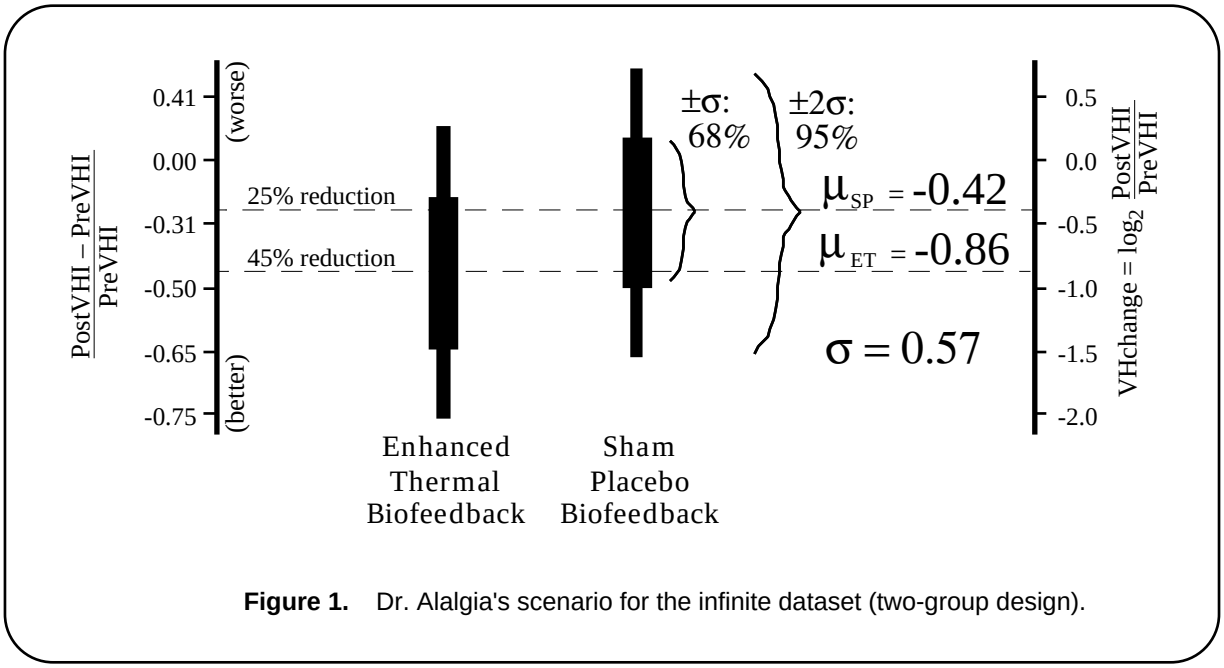

Figure 1. Dr. Alalgia's scenario for the infinite dataset (two-group design).

1. All of these case studies are fictionalized to some degree to make them more useful. I use bogus names for the researchers and drugs, as well as for many of the scientific terms and measurements. This name is derived from **Cephalalgia**, [Gr. *kephalalgia*] head with pain, headache.

Copyright © 1998, Ralph G. O'Brien

At first, the main outcome measure is taken to be the relative change in the VHI values, (PostVHI – PreVHI)/PreVHI. However, the collaborating statistician suggests that a log transform will probably create a better measure for analysis. Logarithms to base 2 are frequently the most convenient, giving the measure

$$\text{VHchange} = \log_2[\text{PostVHI}/\text{PreVHI}] = \log_2\text{PostVHI} - \log_2\text{PreVHI}.$$

VHchange < 0 indicates improvement, that is, a reduction in the VHI.

Volunteer patients will be easier to recruit and retain if they are told that 2/3 will be randomized to the ET arm. Thus an unbalanced design is planned with sample size weights of $w_{ET} = 2/3$ and $w_{SP} = 1/3$, a randomization ratio of 2:1. Here, *recruiting* efficiency outweighs the *statistical* efficiency of a balanced design.

After the design is set, the statistician elicits from Alalgia a vision about the *infinite dataset* that will be sampled. (It is too confusing for many researchers to discuss "population" distributions and parameters, and there is no need to define "effect sizes" directly.) Alalgia has some pilot data, but not nearly enough to use formally in specifying what the means and common standard deviation may be for the infinite dataset. So, based on his knowledge and experience, and his scant pilot data, he conjectures that the ET group's true mean on VHchange is $\mu_{ET} = -0.86$, corresponding to almost a 45% reduction in VHI (because, $\log_2[0.55] = -0.86$), while $\mu_{SP} = -.42$, roughly a 25% reduction in VHI with litosamine plus the sham placebo feedback.

Determining the common standard deviation is more difficult. In response to the statistician's probing, Alalgia postulates that the middle 95% of the ET patients may have VHchange scores between -2.00 and +0.28, corresponding to a VHI reduction of 75% to an increase of 21%. Assuming Normality for VHchange, this defines a 4σ range, making $\sigma = 2.28/4 = 0.57$. Applying $\sigma = 0.57$ to the SP group gives a mid-95% range of -1.56 to +0.72, corresponding to a reduction of 66% to an increase of 65%. This scenario is depicted in Figure 1.

It is prudent to bracket values for σ in order to assess the impact of small differences in this difficult conjecture. Let us use $\sigma = 0.45$ and $\sigma = 0.65$, as well.

Dr. Alalgia estimates that he has enough resources to study N = 21 patients. The GCRC's Scientific Review Committee hopes that a valid study can be conducted by using fewer than the 105 hospital bed-days requested (21 patients $\times$ 5 days/patient). Thus, N= 15 will also be assessed along with a substantial number, N = 33. This study is an early clinical trial of this behavioral therapy, and research in biofeedback has a history of conflicting findings. Thus Alalgia and his statistician feel that a directional ("one-tailed") test is unwarranted at this time; only the nondirectional ("two-tailed") test will be used, but it is interesting to view the one-tailed powers anyway. Finally, Alalgia will concentrate on the powers for $\alpha = .05$, but he is curious to see the

Copyright © 1998, Ralph G. O'Brien

powers under $\alpha = .01$. Under these conditions, will N = 21 provide acceptable power? Would fewer suffice?

UnifyPow input for ordinary 2-group t test

```
mu -.86 -.42            [Will handle G independent groups]
SD .45 .57 .65          [At least one value required]
weight 2 1              [Optional; default = balanced]
alpha .05 .01           [Optional; default = 0.05]
NTotal 15 21 33         [At least one value required]
```

Examine the UnifyPow (O'Brien, 1998) input for this "mu" problem, noticing how directly the user goes from building the design and scenario to the power analysis. Here we obtain powers for Dr. Alalgia's scenario by crossing all combinations of $N \in \{15, 21, 33\}$ with $\sigma \in \{0.45, 0.57, 0.65\}$. The $\alpha = .05$ results for the nondirectional test indicate that N = 21 may be too few subjects to run: the .05-based power is 0.35 for $\sigma = 0.57$. Even for N = 33 and $\sigma = 0.45$, the power is only 0.83. Dr. Alalgia is discouraged about these results and understands that the GCRC cannot afford to provide more than 105 bed-days for his project. He decides to restructure the design, which we handle next.

UnifyPow output for ordinary 2-group t test

```
Scenario: mu -.86 -.42 .     <== optional " ." allows comment to follow
Ordinary t test
```

		Standard Deviation							
		0.45			0.57			0.65	
		Total N			Total N			Total N	
		15	21	33	15	21	33	15	21	33
		Power	Power	Power	Power	Power	Power	Power	Power	Power
Alpha	Type									
0.05	2-tail t	.380	.518	.727	.257	.353	.526	.209	.284	.427
	1-tail t	.519	.652	.828	.379	.485	.655	.318	.407	.559
0.01	2-tail t	.156	.260	.472	.091	.147	.276	.068	.108	.201
	1-tail t	.235	.358	.581	.145	.219	.372	.112	.167	.283

Copyright © 1998, Ralph G. O'Brien

Even researchers with limited experience in statistical planning readily understand the rationale and implications of a power analysis like that given here. The results provided by UnifyPow show concretely how power is increased by (1) increasing the α level, (2) decreasing the variance, (3) increasing the total sample size, and (4) using a directional test.

Copyright © 1998, Ralph G. O'Brien

Matched-pairs t test comparing two correlated means

Dr. Seth Alalgia now considers a new design to study the efficacy of the enhanced thermal (ET) biofeedback therapy relative to the sham-placebo (SP) therapy in the relief of vascular headache. (See previous example.) Patients will be run in *pairs* after being matched on several factors, including a screening version of the Vascular Headache Index (VHI).The 2-day pretreatment phase will be changed so that subjects are only seen as outpatients on Monday and Tuesday to complete the PreVHI measure and begin their standardized medical treatment with litosamine. Each pair will be admitted on Wednesday morning, already randomly split into the two treatment groups. The ET subject will be treated first, getting legitimate biofeedback. Then the SP subject will be treated as a "yoked control," getting the exact same sequence of biofeedback as the ET subject did, in essence, random (non-contingent) biofeedback. The therapy will be repeated on Thursday. The PostVHI measure will be assessed Thursday afternoon and Friday morning just before subjects are discharged. The same outcome measure, VHchange, will be used, so that our paired differences are $D = VHchange_{ET} - VHchange_{SP}$. It is hoped that by matching and yoking pairs of patients, the error variance can be markedly reduced. Note also that now only three GCRC bed-days per subject will be used, rather than five, as before.

Dr. Alalgia believes that this one-group (matched-pairs) study carries with it the same treatment effects and similar variability as for the two-group study. Thus, he takes $[\mu_{ET}, \mu_{SP}] = [-0.86, -0.42]$, so that $\mu_D = -0.44$. Dr. Alalgia believes that the standard deviations will be greater under this design, because the outpatient PreVHI measurements are not as well controlled and PostVHI measurement is taken over fewer hours. Further, he feels that the SP (random biofeedback) patients will have greater variability than the ET patients. He sets $[\sigma_{ET}, \sigma_{SP}] = [0.60, 0.80]$. The correlation between $VHchange_{ET}$ and $VHchange_{SP}$ is suspected to be at least $\rho = 0.50$. It is convenient to display these values in an "SD-Corr" matrix,

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.60 & 0.50 \\ 0.50 & 0.80 \end{bmatrix}$$

In this case, $\sigma_D = [0.60^2 + 0.80^2 - 2(0.50)(0.60)(0.80)]^{1/2} = 0.72$. Thus, this one-group t test is assessing H_0: $\mu_D = 0$ assuming D is Normal with $\mu_D = -0.44$ and $\sigma_D = 0.72$. σ_D can be varied by a multiplicative factor, m, giving the general equation,

$$\sigma_D(\rho, m) = m[\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2]^{1/2}$$

Dr. Alalgia is prepared to study up to 25 pairs of patients, which will require $25 \times 2 \times 2 = 100$ bed-days. Is this enough? Too many? What if the correlation is greater, say $\rho = 0.60$? What if the standard deviations are, say, 20% larger (m = 1.20) than Dr. Alalgia conjectures? Examining the input for this "PairedMu" UnifyPow problem shows how easily one handles such questions. Here we obtain powers for Dr. Alalgia's scenario by crossing all combinations of $N \in \{17, 25\}$ pairs with $\rho \in \{0.50, 0.60\}$, and $m = \in \{1.0, 1.2\}$.

Copyright © 1998, Ralph G. O'Brien

UnifyPow input for matched-pairs t-test example

```
PairedMu -.86 -.42      [Will handle G independent groups x 2 repeated measures]
SD .60 .80              [Requires exactly one pair of values]
corr .50 .60            [At least one value required]
SDMult 1.0  1.2         [Optional; default = 1.0]
alpha .05 .01           [Optional; default = 0.05]
Ntotal 17 25            [At least one value required]
```

UnifyPow output for matched-pairs t-test example

Scenario: PairedMu -.86 -.42 & SD 0.6 0.8
Matched-pairs t test

		x SD (SD Multiplier)					
		1		1.2			
		Corr(Y1, Y2)		Corr(Y1, Y2)			
		0.5	0.6	0.5	0.6		
		Total N	Total N	Total N	Total N		
		17	25	17	25	17	25
		Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Alpha	Test Type								
0.05	2-tail t	.657	.833	.744	.900	.504	.684	.588	.771
	1-tail t	.777	.906	.846	.949	.641	.795	.718	.862
0.01	2-tail t	.375	.604	.469	.715	.245	.418	.312	.518
	1-tail t	.491	.709	.588	.805	.343	.529	.421	.629

Under Dr. Alalgia's scenario for the means and standard deviations, and with $\rho = 0.50$, the $\alpha = .05$, nondirectional test with N = 25 has powers of 0.83 and 0.68 for m = 1.0 and 1.2, respectively. Under $\rho = 0.60$, the corresponding powers are 0.90 and 0.77. Dr. Alalgia decides to go forward with N = 25 and feels he can now defend this choice before the GCRC's review committee.

Copyright © 1998, Ralph G. O'Brien

References

Blanchard EB, Appelbaum KA, Radnitz CL, Morrill B, Michultka D, Kirsch C, Guarnieri P, Hillhouse J, Evans DD, Jaccard J, Barron KD (1990), "A Controlled Evaluation of Thermal Biofeedback and Thermal Biofeedback Combined with Cognitive Therapy in the Treatment of Vascular Headache," *J Consulting Clinical Psychology,* 2, 216-224.

Pocock SJ, Simon R (1975), "Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial," *Biometrics,* 31, 103-115.

O'Brien RG (1998), "A Tour of UnifyPow: A SAS Module/Macro for Sample-Size Analysis," *Proceedings of the 23nd SAS Users Group International Conference*, Cary, NC, SAS Institute, 1346-1355. [For this and all other UnifyPow documentation and freeware, visit http://www.bio.ri.ccf.org/power.html.]

Copyright © 1998, Ralph G. O'Brien