PUMP manuscript draft

Power Under Multiplicity Project (PUMP): An R package for estimating statistical power, minimum detectable effect sizes (MDES's) and sample sizes when adjusting for multiple hypothesis tests Draft outline

TODO list

(written by Kristen)

Topics where I think we need more detail:

- Luke: would you like to add a paragraph on the optimization procedure?
- give more context for people who aren't familiar with multilevel experiments. What is blocking? clustering? Why might you want to choose a particular RCT design?
- say this explicitly and motivate our broad choice: we are fitting mixed effects regression models. We don't explicitly say this anywhere, and perhaps should spend just 1 sentence motivating this.
- effect size: what it is, how to decide what it is
- correlation between test statistics, relationship to correlation between outcomes
- organization of RCT designs in Overview–do we want to re-arrange given our new naming convention?
- explanation of MTP procedures, like in original paper

Optional topics:

- advice on binary outcomes, consider how it affects assumptions
- reference to literature on summarizing/collapsing outcomes?

Introduction

Overview

Researchers are often interested in testing the effectiveness of an intervention on multiple outcomes, for multiple subgroups, at multiple points in time, or across multiple treatment groups. The resulting multiplicity of statistical hypothesis tests can lead to spurious findings of effects. Multiple testing procedures (MTPs) are statistical procedures that counteract this problem by adjusting p-values for effect estimates; generally, p-values are adjusted upward to require a higher burden of proof. When not using an MTP, the probability of finding false positives increases, sometimes dramatically, with the number of tests. When using an MTP, this probability is reduced.

However, an important consequence of MTPs is a change in statistical power that can be substantial. That is, the use of MTPs changes the probability of detecting effects when they truly exist, compared with the situation when the multiplicity problem is ignored. Unfortunately, while researchers are increasingly using MTPs, they frequently ignore the power implications of their use when designing studies. Consequently, in some cases sample sizes may be too small, and studies may be underpowered to detect effects as small as a desired size. In other cases, sample sizes may be larger than needed, or studies may be powered to detect smaller effects than anticipated.

Researchers typically worry that moving from one to multiple hypothesis tests and thus employing MTPs results in a loss of power. However, that need not always be the case. Power is indeed lost if one focuses on individual power — the probability of detecting an effect of a particular size or larger for each particular hypothesis test, given that the effect truly exists. However, in studies with multiplicity, alternative definitions

of power exist and in some cases may be more appropriate (Chen (2011); Dudoit (2003); Senn and Bretz (2007); Westfall, Tobias, & Wolfinger, 2011).

For example, when testing for effects on multiple outcomes, one might consider 1-minimal power: the probability of detecting effects of at least a particular size (which can vary by outcome) on at least one outcome. Similarly, one might consider 1/2-minimal power: the probability of detecting effects of at least a particular size on at least 1/2 of the outcomes. Also, one might consider complete power: the power to detect effects of at least a particular size on all outcomes. The choice of definition of power depends on the objectives of the study and on how the success of the intervention is defined. The choice of definition also affects the overall extent of power.

The methodological developments implemented in our R package, PUMP, and described in this paper are focused on the multiplicity problem that arises in randomized control trials (RCT's) that test an intervention's impacts on a modest number of outcomes. For example, in education a researcher might design a trial to investigate the effects of a mentoring program on three outcomes related to social and emotional development — measures of social competence, emotional competence and self-regulation. For this type of research, the literature and tools for estimating statistical power, or for estimating sample size requirements of minimum detectable effect sizes (MDES's) that achieve a desired level of statistical power, are extensive.[1] However, to our knowledge, no tools exist that take multiplicity into account.

The PUMP package fills this gap. It allows users to estimate statistical power, sample size requirements or MDES's for multiple definitions of statistical power, when applying any of five common MTPs — Bonferroni, Holm, single-step and step-down versions of Westfall-Young, and Benjamini-Hochberg and when using any of the following RCT designs and estimation models:

- Individual random assignment
- Blocked individual random assignment
  - Constant effects model (2 levels - e.g., students blocked within teachers)
  - Fixed effects model (2 levels and 3 levels - e.g., students blocked within teachers within schools)
  - Random effects model (2 levels, 3 levels)
- Cluster random assignment
  - 2-level model (treatment at Level 2 - e.g., at teacher level)
  - 3-level model (treatment at Level 3 - e.g., at school level)
- Blocked cluster random assignment
  - 3-level fixed-effects model (treatment at Level 2)
  - 3-level random-effects model (treatment at Level 2)

For details about the assumptions for each of the estimation models, see TODO.

Review of the multiple testing problem in a frequentist framework

This paper focuses on the frequentist framework of hypothesis testing, as it is currently the prevailing framework in education and other social policy research.

In the frequentist framework, when framing impacts in terms of effect sizes, one typically tests a null hypothesis of no effect, $H_{0_m} : ES_m = 0$, against an alternative hypothesis $H_{1_m} : ES_m \neq 0$ for a two-sided tests or $H_{1_m} : ES_m > 0$ or $H1_m : ES_m < 0$ for a one-sided test. For the purposes of computing power researchers specify an alternative hypothesis of at least a particular effect size. A significance test, such as a two-sided or one-sided t-test, is then conducted, and one obtains a test statistic given by

$$t_m = \frac{\hat{ES}_m}{SE(\hat{ES}_m)}, \tag{1}$$

from which a raw p-value is computed. Here, the term "raw" is used to distinguish this p-value from a p-value that has been adjusted for multiple hypothesis tests, as discussed below. The raw p-value is the probability

[1] For example, power estimating tools frequently used in social science research include Dong and Maynard (2013); Hedges and Rhoads (2010); Raudenbush et al. (2011); Spybrook et al. (2011)).

of a test statistic being at least as extreme as the one observed, given that the null hypothesis is true. For a two-sided test, which is the focus of the discussion going forward (although the PUMP package also allows for one-sided tests), the raw p-value for test m is $p_m = 2 * Pr(T_m \geq |t_m|)$.[2] This expression means we use our knowledge of the sampling distribution of the t-statistic, and we identify where our observed test statistic falls in that distribution when it is centered around zero.

When testing a *single* hypothesis under this framework (effects are being assessed on just one outcome, so that $M = 1$), researchers typically specify an acceptable maximum probability of making a Type I error, α. A Type I error is the probability of erroneously rejecting the null hypothesis when it is true. The quantity α is also referred to as the significance level. If $\alpha = 0.05$, then the null hypothesis is rejected if the p-value is less than 0.05, and it is concluded that the intervention had an effect because there is less than a 5% chance that this finding is a false positive.

When one tests *multiple* hypotheses under this framework (such that $M > 1$) and one conducts a separate test for each of the hypotheses with $\alpha = 0.05$, there is a *greater* than 5% chance of a false positive finding in the study. If the multiple tests are independent, the probability that at least one of the null hypothesis tests will be erroneously rejected is 1-Pr(none of the null hypotheses will be erroneously rejected) $= 1-(1-\alpha)^M$. Therefore, if researchers are estimating effects on three outcomes, and if these outcomes are assumed independent, the probability of at least one false positive finding is 0.14. If the researchers were instead estimating effects on five independent outcomes, the probability of at least one false positive finding is 0.23. This Type I error inflation for independent outcomes demonstrates the crux of the multiple testing problem. In practice, however, the multiple outcomes are usually at least somewhat correlated, which makes the test statistics correlated and reduces the extent of Type I error inflation. Nonetheless, any error inflation can still make it problematic to draw reliable conclusions about the existence of effects. As introduced above, to counteract the multiple testing problem, MTPs adjust p-values upward.[3] The following paragraphs describe how using a multiple testing procedure protects against false positives.

Recall that the power of an individual hypothesis test is the probability of rejecting a false null hypothesis of at least a specified size. If raw p-values are adjusted upward, one is less likely to reject the null hypotheses that are true (meaning there is truly no effect of at least a specified size), which reduces the probability of Type I errors, or false positive findings. Reducing this probability is the goal of MTPs. However, if raw p-values are adjusted upward, one is also less likely to reject the null hypotheses that are false (meaning there truly is an effect of at least a specified size). Therefore, all MTPs reduce individual power (the power of separate hypothesis tests for each outcome) compared with the situation when no multiplicity adjustments are made or the situation when there is only one hypothesis test.

MTPs also reduce all other definitions of power compared with the situation when no multiplicity adjustments are made – but not necessarily compared with the situation when there is only one hypothesis test. For example, 1-minimal power, the probability of detecting effects (of at least a specified size) on at least one outcome – after adjusting for multiplicity – is typically greater than the probability of detecting an effect of the same size on a single outcome. This increase may or may not occur with other definitions of power (e.g., the probability of detecting a third, half, or all false null hypotheses).

Using MTPs to protect against spurious impact findings

The MTPs that are the focus of this paper fall into two different classes. The first class reframes Type I error as a rate across the entire set or "family" of multiple hypothesis tests. This rate is called the familywise error rate (FWER; RN33098). The FWER is typically set to the same value as the probability of a Type I error for a single test, or to α. MTPs that control the FWER at 5% adjust p-values in a way that ensures that the probability of at least one Type I error across the entire set of hypothesis tests is no more than 5%. The MTPs introduced by Bonferroni (Dunn (1959), Dunn (1961)), Holm (1979), and Westfall and Young (1993) control the FWER.

[2] For a one-sided test, depending on the direction of our alternative hypothesis, the raw p-value for test m is computed as $p_m = Pr(T_m \leq t_m)$ or $p_m = Pr(T_m \geq t_m)$.

[3] Alternatively, MTPs can decrease the critical values for rejecting hypothesis tests. For ease of presentation, this paper focuses only on the approach of increasing p-values.

The second class of MTPs takes an entirely different approach to the multiple testing problem. MTPs in this class control the false discovery rate (FDR). Introduced by Benjamini and Hochberg (1995), the FDR is the expected proportion of all rejected hypotheses that are erroneously rejected. The two-by-two representation in Table 1 is often found in articles on multiple hypothesis testing, and helps to illustrate the difference between FWER and FDR. Let M be the total number of tests. Therefore, we have M unobserved truths: whether or not each null hypothesis is true or false. We also have M observed decisions: whether or not the null hypotheses were rejected, because the p-values were less than α. In Table 1, A, B, C and D are four possible scenarios: the numbers of true or false hypotheses not rejected or rejected. $M0$ and $M1$ are the unobservable numbers of true null and false null hypotheses. R is the number of null hypotheses that were rejected, and $M - R$ is the number of null hypotheses that were not rejected.

In Table 1, B is the number of erroneously rejected null hypotheses, or the number of false positive findings. Therefore, the FWER is equivalent to $Pr(B > 0)$, the probability of at least one false positive finding. Recall the examples above about Type I error inflation when testing for effects on independent outcomes in the case that α is set to 0.05 and no MTPs are applied. The Type I error was almost 10% when testing effects on two independent outcomes and 23% when testing effects on five independent outcomes. These Type I error rates both correspond to the FWER. The goal of MTPs that control the FWER is to bring these percentages back down to 5%.

Also in Table 1, the FDR is equal to $E(\frac{B}{R})$ but is defined to be 0 when $R = 0$, or when no hypotheses are rejected. As is frequently noted in the literature (e.g., Shaffer (1995); Schochet (2008)), the FWER and FDR have different objectives. Control of the FWER protects researchers from spurious findings and so may be preferred when even a single false positive could lead to the wrong conclusion about the effectiveness of an intervention. On the other hand, the FDR is more lenient with false positives. Researchers may be willing to accept a few false positives, B, when the total number of rejected hypotheses, R, is large. Note that under the complete null hypothesis that all M null hypotheses are null, the FDR is equal to the FWER, because when referring back to Table 1 we have $FWER = P(R > 0) = E(\frac{B}{R}) = FDR$. However, if any effects truly exist, then $FWER \geq FDR$.

As a result, in the case where there is at least one false null hypothesis (at least one true effect at least as large as a specified effect size), an MTP that controls the FDR at 5% will have a Type I error rate that is greater than 5%. Note that MTPs may provide either weak or strong control of the error rate they target. An MTP provides weak control of the FWER or the FDR at level α if the control can only be guaranteed when all nulls are true, or when the effects on all outcomes are zero. An MTP provides strong control of the FWER or FDR at level if the control is guaranteed when some null hypotheses are true and some are false, or when there may be effects on at least some outcomes. Of course, strong control is preferred.

The list of multiple testing procedures supported by the PUMP package can be found below.

Method	Comment
None	No adjustment
Bonferroni	The classic (and conservative) multiple testing correction
Holm	Step down version of Bonferroni
BH	Benjamini-Hochberg
WY-SS	Westfall-Young, Single Step
WY-SD	Westfall-Young, Step Down

Estimating power, MDES and sample size in studies with multiple outcomes

Power estimation strategy

We take an innovative simulation-based approach to estimating power for multiple outcomes, as introduced in Porter (2018). In order to estimate power for a single outcome, we can often use closed-form algebraic expressions, which are derived from the assumed model. However, with multiple outcomes, finding such expressions can be quite difficult, or even impossible depending on the multiple testing procedure. In cases

where it is possible to find a closed-form expression, we would need to find expressions for every design, MTP, and definition of power. Importantly, we would also need to find new expressions for each number of outcomes, which quickly becomes an intractable problem! In some cases, such as permutation-based procedures like Westfall-Young approaches, a closed-formed solution does not exist. Instead, we rely on simulation to calculate estimated power. The approach outlined below can estimate power for any scenario.

If we were to rely on a full simulation approach, we could use the following method to estimate power:

1. *Simulate a data sample according to the joint alternative hypothesis.* First, we formulate what we will refer to as the *joint alternative hypothesis*, which is the set of outcomes we assume to have nonzero treatment effects. We define ψ_m to be the treatment effect for outcome m, with M total outcomes. If we have $M = 3$ treatments, one possible joint alternative hypothesis is that all outcomes have nonzero effects: $H_A : \psi_1 \neq 0, \psi_2 \neq 0, \psi_3 \neq 0$. Another possible joint alternative hypothesis is that only the first two outcomes have nonzero effects: $H_A : \psi_1 \neq 0, \psi_2 \neq 0, \psi_3 = 0$. Then, we would generate simulated data under the joint alternative hypothesis.
2. *Calculate test statistics t_1 under the joint alternative hypothesis.* Given simulated data, for example we could fit M regression models (specified to match model assumptions).
3. *Calculate adjusted p-values.* The test statistics can be used to compute raw (unadjusted) p-values.
4. *Repeat above steps (1 through 3) for a large number of iterations.* Denote the number of iterations *tnum*. Repeating steps 1-3 *tnum* times results in a matrix of unadjusted p-values which we call **F**, and is of dimension $tnum \times M$.
5. *Adjust p-values.* For each row, corresponding to one simulated dataset, the M raw p-values corresponding to the M hypothesis tests can be adjusted according to the desired multiple testing procedure to generate a new matrix **G**. For Bonferroni, Holm, and BH adjustments, we use the function `p.adjust` in R (found in the `stats` package). We developed our own function for implementing adjustment using the Westfall-Young procedures.
6. *Calculate hypothesis rejection indicators.* For each MTP, the matrix of adjusted p-values **G** can then be compared with a specified value of α (the default is 0.05, but the value can be changed by the user). For each row, which is one iteration of simulated data, we record whether or not the null hypothesis was rejected for each outcome. This process results in a new matrix **H**, which contains hypothesis rejection indicators, and is still of dimension $tnum \times M$. Using **H**, we can compute all definitions of power.
7. *Calculate power.* To compute the different definitions of power:
   - *Individual power for outcome m* is the proportion of the `tnum` rows in which the null hypothesis m was rejected (the mean of column m of **H**).
   - *d-minimal power* is the proportion of the *tnum* rows in which at least d of the M null hypotheses were rejected.[4]
   - *Complete power* is the proportion of the *tnum* rows in which all of the null hypotheses were rejected based on the raw p-values rather than adjusted p-values.

The reason that complete power is based on raw p-values is that the probability of all tests having a raw p-value less than 0.05 when the null hypothesis is true is less than the probability that any single test would have a p-value less than α by chance (Koch and Gansky (1996); Westfall et al., 2011).[5]

Above, we outline a full simulation-based approach for calculating power. We can simplify this process by skipping the first step. Given an assumed model and correlation structure for the test statistics, we can directly sample from $f(t_1)$, the joint alternative distribution of the test statistics. This shortcut vastly improves both the simplicity and the speed of computation. In summary, our approach is:

1. **Sample** *test statistics t_1 under the joint alternative hypothesis.*
2. *Calculate adjusted p-values.*
3. *Repeat above steps (1 through 2) for a large number of iterations and calculate power.*

[4]Note that others refer to 1-minimal power simply as "minimal power" (e.g., Maurer and Mellein (1988); Chen (2011); Westfall, Tobias, & Wolfinger, 2011), "disjunctive power" (e.g., Bretz, Hothorn, & Westfall, 2011), or "any pair" power (Ramsey (1978)). Chen (2011) use the terminology of "r-power" for what is refered to here as d-minimal power for d>1.

[5]Complete power does not in itself require unadjusted tests. The above approach for not adjusting tests assumes that all tests must to be statistically significant in order to claim impacts on all outcomes.

We now describe how to sample from $f(t_1)$ directly. First, we assume a particular research design and model. Define ψ_m as the treatment effect for outcome m. Then, we can also express the treatment effect in terms of effect size:

$$ES_m = \frac{\psi_m}{\sigma_m}$$

where σ_m is the standard deviation of outcome Y_m. In order to calculate power, we are interested in the standard error of the estimated effect size, which we denote as

$$Q_m = SE(\hat{ES}_m).$$

The quantity Q_m is defined by the assumed model, and can be a function of the number of units at different levels, the percent of units treated, the assumed R^2, and other parameters. When analyzing actual data, we do not know the true value of Q_m, so we would need to estimate Q_m by plugging in either known or estimated values of the relevant parameters. Some parameters, such as the percent of units treated, are known, while others, such as the R^2 at different levels, would need to be estimated.

Given an estimate of $\hat{Q}_m$, we can arrive at the distribution of test statistics. When testing the hypothesis for outcome m, the test statistics for a t test is:

$$t_m = \frac{\hat{ES}_m}{\hat{Q}_m}$$

with degrees of freedom df, also defined by the assumed model. Under the alternative hypothesis for outcome m, t_m has a t distribution with degrees of freedom df and mean $\hat{ES}_m/\hat{Q}_m$. Finally, we choose the correlation matrix between test statistics ρ to sample from the joint distribution of $t_m, m = 1, \ldots M$.

From the power formulas, we can then also arrive at MDES and sample size calculations. From Dong and Maynard (2013), in general the MDES can be estimated as

$$MDES = MT_{df} \times SE/\sigma_m$$

where MT_{df} is known as the multiplier and is the sum of two t statistics with degrees of freedom df. For one-tailed tests, $MT_{df} = t^{\star}_{\alpha} + t^{\star}_{1-\beta}$ where α is the type I error rate and β is the desired power. For two-tailed tests, $MT_{df} = t^{\star}_{\alpha/2} + t^{\star}_{1-\beta}$. We do not explain the details of the derivations here; for more details and understanding, see Dong and Maynard (2013) or Bloom (2006). Manipulating this expression then results in sample size formulae.

The p-value adjustment using Westfall-Young procedures is the most complex. We briefly outline the algorithm below. Similar to above, we first explain a full simulation approach, and then discuss our simplification. Under a full simulation approach, we would first generate a single dataset under the joint alternative hypothesis and calculate a set of observed test statistics. Then, we would resample the simulated data, say $B = 3,000$ times, under the joint null hypothesis, and calculate test statistics on each of these resampled datasets to generate a distribution of test statistics under the joint null distribution. Next, we compare the distribution of observed test statistics to the distribution of test statistics under the joint null distribution to calculate p-values. We would then re-generate a new simulated dataset, and repeat the process. If we were to generate $tnum = 10,000$ datasets under the joint alternative hypothesis, for each of these datasets we also generate $B = 3,000$ resampled datasets under the joint null, so we would have to generate $10,000 \times 3,000$ datasets!

When we skip the simulation step, for each iteration t in $1, \ldots, tnum$ we generate a set of observed test statistics from the joint alternative distribution. Then, we augment the method by drawing B samples of test statistics under the joint null rather than resampling the data B times. Under the null hypothesis, t_m has a t distribution with degrees of freedom df and mean 0. As before, we then compare the distribution of observed test statistics to the distribution of test statistics under the joint null distribution to calculate p-values. Westfall-Young procedures are computationally intensive, so the approach of skipping the simulated data step is particularly helpful here. This approach substantially reduces computational time by drawing test statistics directly rather than resampling data. Additionally, we find that sampling the test statistics

can actually produce more accurate results in some cases for WY procedures; see TODO appendix for more details.

Note that this approach of simulating test statistics builds on work by Bang (2005), who use simulated test statistics to identify critical values based on the distribution of the maximum test statistics. Their approach produces the same estimates as the approach described here for the single-step Westfall-Young MTP. Chen (2011) derived explicit formulas for $d-$minimal powers of stepwise procedures and for complete power of single-step procedures, but only for 1, 2, or 3 tests. The approach presented here is more generally applicable, as it can be used for all MTPs, for any number of tests, and for all definitions of power discussed in the present paper.

Randomized Control Trial Designs and Models

When designing a study, the researcher has two main choices. First, the researcher chooses the design of the experiment, including the number of levels, and at which level randomization occurs. Second, and separately, the researchers choose an assumed model, including whether intercepts and treatment effects should be treated as constant, fixed, or random. For the same experimental design, the analyst can make sometimes choose from a variety of possible models, and these two decisions should be conceptually separated from each other.

For the design, the PUMP package supports designs with 1, 2, or 3 levels, with randomization occurring at any level. For example, a design with 2 levels and randomization at level 1 is a blocked design. A design with 3 levels and randomization at level 3 is a cluster design. For modeling, we have the following choices.

- Whether level 2 and level 3 intercepts are:
  - fixed: intercepts are fixed effects constrained to have mean 0.
  - random: intercepts are considered to be Normally distributed, allowing for partial pooling.
- Whether level 2 and level 3 treatment effects are:
  - constant: all units in a level have the same single average impact.
  - fixed: each unit within a level has an individual estimated impact, with an additional mean impact.
  - random: treatment impacts are Normally distributed around a mean impact.

The research design is denoted by d, followed by the number of levels and randomization level, so `d3.1` is a 3-level design with randomization at level 1. The model is denoted by m, followed by the level and the assumption for the intercept, either f or r and then the assumption for the treatment impacts, c, f, or r. For example, `m3ff2rc` means at level 3, we assume fixed intercepts and treatment impacts, and at level 2 we assume random intercepts and constant treatment impacts. The full design and model are specified by concatenating these together, e.g. `d3.2_m3ff2rc`.

The full list of supported models is below. We also including the corresponding names from the PowerUP! package where appropriate. For more details about each model, see the appendix.

Code	Design	Model	PowerUp
d1.1_m2cc	d1.1	m2cc	n/a
d2.1_m2fc	d2.1	m2fc	blocked_i1_2c
d2.1_m2ff	d2.1	m2ff	blocked_i1_2f
d2.1_m2fr	d2.1	m2fr	blocked_i1_2r
d2.2_m2rc	d2.2	m2rc	simple_2c_2r
d3.1_m3rr2rr	d3.1	m3rr2rr	blocked_i1_3r
d3.2_m3ff2rc	d3.2	m3ff2rc	blocked_c2_3f
d3.2_m3rr2rc	d3.2	m3rr2rc	blocked_c2_3r
d3.3_m3rc2rc	d3.3	m3rc2rc	simple_c3_3r

Understanding design parameters

The table below shows the parameters that influence Q_m formulae for different designs.

Parameter	Description
nbar	the harmonic mean of the number of level 1 units per level 2 unit (students per school)
J	the number of level 2 units (schools)
K	the number of level 3 units (district)
Tbar	the proportion of units that are assigned to the treatment
numCovar.1	number of Level 1 (individual) covariates
numCovar.2	number of Level 2 (school) covariates
numCovar.3	number of Level 3 (district) covariates
R2.1	percent of variation explained by Level 1 covariates
R2.2	percent of variation explained by Level 2 covariates
R2.3	percent of variation explained by Level 3 covariates
ICC.2	level 2 intraclass correlation
ICC.3	level 3 intraclass correlation
omega.2	ratio of variance of level 2 average impacts to variance of level 2 random intercepts
omega.3	ratio of variance of level 3 average impacts to variance of level 3 random intercepts

A few parameters warrant more explanation. The quantity ICC is the Intraclass Correlation, and gives a measure of variation at different levels of the model. For each outcome, the ICC for each level is defined as the ratio of the variance at that level divided by the overall variance of the individual outcomes. The ICC is for the unconditional model, and therefore include the variation due to covariates. For each outcome, the quantity ω for each level is the ratio between impact variation at that level and mean variation at that level. The R^2 expressions are the percent of variation at a particular level predicted by covariates specific to that level.

Estimating MDES and sample size

Frequently, a researcher's main concern with power is prospectively calculating either the minimum detectable effect size (MDES) from a possible study, or determining the necessary sample size. Given our simulation approach, it is necessary to perform a search algorithm to calculate MDES and sample size. The user provides a particular target power, say 80%. To perform a search, we calculate power over a range of different MDES or sample size values, and then find the value (of sample size of MDES) that is within a specified tolerance of the target power. We discuss the algorithm for sample size, although the approach for MDES is the same.

First, we begin by bounding the possible range of sample size values. Bonferroni is the most conservative correction, so it would result in the largest possible sample size, and thus a Bonferroni correction provides our upper bound. If we are interested in complete power, we have a larger upper bound; in order to have complete power of 0.8, we would need each outcome to have an individual power of $0.8^{(1/M)}$. On the other hand, our least conservative correction is doing no correction at all, and would result in the smallest possible sample size, so no adjustment provides our lower bound. If we are interested in minimal power, we must have a smaller lower bound; in order to have 1-minimal power of 0.8, each outcome needs to have individual power of $1 - (1 - 0.8)^{(1/M)}$.

Second, given our lower and upper bounds, we now use optimization to find the sample size. **TODO** Describe the optimization procedure. (Luke?)

Finally, given a proposed sample size value by the optimization procedure, we calculate the power for the given power to ensure it is within a specified tolerance of the target power. The default tolerance is 1, so given a target power of 80, we check whether the given sample size is between 79 and 81.

Validation

We completed extensive validation checks to ensure our power calculation procedure was correct. First, we compared our power estimates in scenarios with only one outcome, $M = 1$, to PowerUpR!. Without a multiple testing procedure adjustment, our estimates match. Second, in order to validate our estimates under multiplicity, we used a simulation approach. We generated many iterations of data according to the assumed design and model, calculated p-values, and calculated an empirical estimate of power. Finally, using

a binomial distribution we construct confidence intervals for the power estimate. Then, we validated that the PUMP estimates fall within these confidence intervals.

A more detailed explanation of the validation procedure can be found in the Appendix, and full validation code and results are in our github repository **TODO**. For some scenarios, we have discrepancies from PowerUp resulting from different modeling choices. For example, for certain models PowerUp assumes the intraclass correlation is zero, while we allow for nonzero values. When there are discrepancies, these are noted in the appendix.

PUM-P Package (Luke - first October 1)

- Overview of power, mdes, sample size and grid functions + plotting function
- Example(s) - LUKE'S VIGNETTE

Guidance for practice

- Reflections on issues that come up in vignette
  - e.g., how to come up with correlation assumption - what's needed by future researchers so can come to our tool with good assumptions
  - e.g., reinformce importance of looking at ranges when it matters a lot

Appendices

- specifications/derivations for all designs
- how data were generated for simulations
- something about validation maybe? (one example or template?)

References

Bang, Jung, H. 2005. "Sample Size Calculation for Simulation-Based Mulitple-Testing Procedures." Journal Article. *Journal of Biopharmaceutical Statistics* 15: 957–67.

Bloom, Howard S. 2006. "The Core Analytics of Randomized Experiments for Social Research." Government Document. MDRC.

Chen, Luo, J. 2011. "On Power and Sample Size Computation for Multiple Testing Procedures." Journal Article. *Computational Statistics and Data Analysis* 55: 110–22.

Dong, Nianbo, and Rebecca Maynard. 2013. "PowerUP!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies." Journal Article. *Journal of Research on Educational Effectiveness* 6 (1): 24–67.

Dudoit, Shaffer, S. 2003. "Multiple Hypothesis Testing in Microarray Experiments." Journal Article. *Statistical Science* 18 (1): 71–103.

Dunn, Olive Jean. 1959. "Estimation of the Medians for Dependent Variables." Journal Article, 192–97. https://doi.org/10.1214/aoms/1177706374.

———. 1961. "Multiple Comparisons Among Means." Journal Article. *Journal of the American Statistical Association* 56 (293): 52–64. https://doi.org/10.1080/01621459.1961.10482090.

Hedges, Larry V., and Christopher Rhoads. 2010. "Statistical Power Analysis in Education Research." Report. National Center for Special Education Research. http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED509387.

Holm, S. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." Journal Article. *Scand. J. Statist.* 6 (2): 65–70.

Koch, G. G., and M. S. Gansky. 1996. "Statistical Considerations for Multiplicity in Confirmatory Protocols." Journal Article. *Drug Information Journal* 30: 523–33.

Maurer, W., and B. Mellein. 1988. "One New Multiple Test Procedures Based on Independent p-Values and the Assessemtn of Their Powers." Book Section. In *Multiple Hypotheses Testing*, edited by G. Hommel P. bauer and E. Sonnermann, 48–66. Heidelberg, Springer.

Porter, Kristin E. 2018. "Statistical Power in Evaluations That Investigate Effects on Multiple Outcomes: A Guide for Researchers." Journal Article. *Journal of Research on Educational Effectiveness* 11: 267–95.

Ramsey, P. H. 1978. "Power Differences Between Pairwise Multiple Comparisons." Journal Article. *Journal of American Statistical Association* 75: 479–87.

Raudenbush, S. W., J. Spybrook, R. Congdon, X. Liu, A. Martinez, H. Bloom, and C Hill. 2011. "Optimal Design Plus Empirical Evidence (Version 3.0)." Report. http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od.

Schochet, Peter Z. 2008. "Guidelines for Multiple Testing in Impact Evaluations of Educational Interventions. Final Report." Report. Mathematica Policy Research, Inc. P.O. Box 2393, Princeton, NJ 08543-2393. Tel: 609-799-3535; Fax: 609-799-0005; e-mail: info@mathematica-mpr.com; Web site: http://www.mathematica-mpr.com/publications/. http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED502199.

Senn, Stephen, and Frank Bretz. 2007. "Power and Sample Size When Multiple Endpoints Are Considered." Journal Article. *Pharmaceutical Statistics* 6: 161–70. https://doi.org/10.1002/pst.301.

Shaffer, Juliet Popper. 1995. "Multiple Hypothesis Testing." Journal Article. *Annual Review of Psychology* 46 (1): 561–84.

Spybrook, Jessica, H. S. Bloom, Richard Congdon, Carolyn J. Hill, Andres Martinez, and Stephen W. Raudenbush. 2011. "Optimal Design Plus Empirical Evidence: Documentation for the 'Optimal Design' Software Version 3.0." Report. http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od.

Westfall, Peter H, and S Stanley Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Book. Vol. 279. John Wiley & Sons.