**Chapter 3: Proposals, power and sample size estimation**

**Example:** Mothers will be randomized to standard or low phytate maize. Zinc is an essential micronutrient that is often lacking in infant diets in developing countries, and phytate inhibits zinc absorption. Does low phytate maize improve babies’ health? How many subjects are needed? What is the best design?

**3.1 General principles**

**Purposes of power analysis:**

* To get grants
* To assess whether a future study is likely to be “successful”

**Most common situations (may be simplified versions, or parts of more complex designs):**

* Two independent samples, numerical outcome
* Two paired samples, numerical outcome
* Two independent samples, binary outcome

**Quantities involved:**

* Sample size (after drop-out)
* Detectable effect
  + True effect assumed to exist in nature.
  + Large enough to be clinically important.
  + Small enough to be reasonably achievable.
* Power = fraction of studies that would have p<0.05 if the effect specified above actually exists, i.e. Pr( reject H0 | H1 is true )

Can specify any 2 of these and calculate the 3rd.

Ideally the investigator specifies the detectable effect, then we calculate the sample size needed to give adequate (80% or 90%) power.

Often investigators want to see what sorts of effects they can detect given the sample size they can afford, then we may calculate detectable effect to give adequate power.

**ALSO need:**

* For independent numerical outcomes, need \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (see below)
* For paired numerical outcomes, need \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (see below)
* For independent binary outcomes, need \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (see below)

**These can be very hard to get with much confidence.**

**My basic principle: Keep it simple! There is often not sufficient previous information to support elaborate power analyses. (If there is, it should be used.)**

* 1. **Simulation showing what power means.**

Suppose we want to determine power to compare means of two groups, each with N=6 subjects, using two-sided level 0.05 t-tests. The true means are 4 and 5 and the true standard deviations are both 1.

Simulation: Generate lots of datasets (i.e. lots of investigators do the same study), do t-test for each.

Power = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



**Results:** P-values for 100 datasets

0.099 0.053 0.043 0.000 0.098 0.022 0.359 0.022 0.018 0.128 0.104 0.007 0.032 0.011 0.036 0.060 0.069 0.287 0.000 0.056 0.061 0.089 0.013 0.111 0.412 0.018 0.074 0.154 0.436 0.078 0.410 0.002 0.015 0.552 0.634 0.900 0.005 0.051 0.058 0.618 0.134 0.042 0.063 0.041 0.105 0.011 0.007 0.500 0.054 0.774 0.945 0.580 0.045 0.088 0.368 0.562 0.276 0.139 0.301 0.028 0.252 0.034 0.940 0.174 0.009 0.073 0.309 0.979 0.164 0.243 0.365 0.055 0.008 0.101 0.052 0.153 0.012 0.006 0.183 0.639 0.056 0.032 0.068 0.007 0.008 0.119 0.017 0.442 0.027 0.015 0.073 0.316 0.407 0.819 0.064 0.174 0.221 0.260 0.447 0.174

32 are < 0.05 so power ≈ 32%

(364 were <0.05 for 1000 datasets so a better approximation of power ≈ 36.4%)

Using PASS software with SD=1, detectable difference = 1, N=6 per group gives

**Power N1 N2 Ratio Alpha Beta Mean1 Mean2 S1 S2**

0.35742 6 6 1.000 0.05000 0.64258 0.0 1.0 1.0 1.0 **3.3 Estimating population standard deviation (SD)**

This is probably the greatest source of error in sample size calculations, and also usually takes the most effort and care.

**Method 1:** Calculate SD from a previous sample of “similar” subjects in a “similar” situation. Try for “similar” outcomes, ages, conditions, time scales, species, etc.

**Method 2:** Get SD from a previously published paper using “similar” subjects in a “similar” situation. Try for “similar” outcomes, ages, conditions, time scales, species, etc.

**Method 3:** Get SD by asking investigators. “About what would be the lowest and highest values of the outcome?”

For roughly bell-shaped distributions, SD ≈ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

since for a normal distribution, mean ± 2 SD includes about 95% of subjects.

**NOTE 1:** Use SD, not SE.

**NOTE:** Need SD of the outcome being considered, e.g. repeated measurements, proportions, etc..

**Most common situations (may be simplified versions, or parts of more complex designs):**

* Two independent samples, numerical outcome

Need cross-sectional SD of individuals in a sample, or pooled across multiple samples.

* Two paired samples, numerical outcome

Need SD of difference between two measurements on a subject. Two ways to obtain this:

1. From a data that have two measurements on the same subject.

2. From independent groups and the intra-class correlation coefficient, as follows:

Using some simple equations, the variance of the difference in two measurements is

V = Var(Y1 – Y2) = 2 σ2 for independent measurements

V = Var(Y1 – Y2) = 2 σ2 (1-ρ) for paired measurements

where σ2 is the cross-sectional variance in a single group.

Looking at the equations in, say, Rosner, we see that in general for any comparison,

N = [(z1-α/2 + z1-β)/Δ]2×V .

So 100×ρ represents the percent decrease in sample size needed with a paired design compared with the independent sample design.

* Two independent samples, binary outcome

Need true proportion in one of the groups.

**3.4 The process**

**Information that must be determined:**

* Design features
  + Independent or repeated measurements, groups, covariates, etc.
  + Outcome, especially type (numerical, binary, …)
  + Comparisons of interest (e.g. difference, interaction, …)
* Numerical values
  + Difference to be detected or proposed sample size
  + Appropriate SD (or proportion if binary) for comparison of interest

**My steps:**

1. Carefully define the question (outcome, comparisons, etc.) as a hypothesis. See Chapter 1A.
2. Determine a good statistical analysis for the future data.
3. Simplify the analysis if a full power analysis isn’t viable (e.g. some preliminary values not known).
4. Obtain necessary quantities (e.g., SDs, clinical differences).
5. Calculate power, sample size or detectable difference, often for a range of the others.
6. Get investigator response.
7. Return to 1-6 above as necessary.
8. Write it in grant proposal format.

**Final comments (my opinions):**

* Statistical power and sample size analysis cannot generally give very precise results.
* Power analyses sometimes don’t really give much information at all, particularly when little is known about the future situation (eg, standard deviations, detectable differences). Then investigators need to rely on other scientific/biological arguments to decide if the study has a reasonable chance of success. (They may still need a formal sample size calculation for a proposal.)
* Keep it simple.

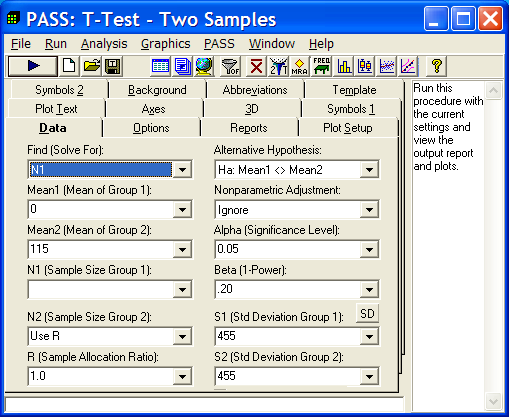
**3.5 Case study:** There are several ways of studying the question “Does low phytate maize improve babies’ health?”, including

* 1. How many mothers needed to see an increase in mean birth weight?
  2. How many mothers needed to see a decrease in LBW incidence?

**Note:** Two common cases: numerical and binary (yes/no) outcomes

**a) Mean birth weight:** Using two-sided level 0.05 t-tests, in order to achieve \_\_\_\_% power to detect a difference of \_\_\_\_g in mean birth weight between low phytate and control maize groups we require \_\_\_\_ subjects per group. Assuming 15% drop-out, we will enroll 300 mothers per group.

PASS (Power and Sample Size) software does these calculations:



**Allocation**

**Power N1 N2 Ratio Alpha Beta Mean1 Mean2 S1 S2**

0.80043 246 246 1.000 0.05000 0.19957 0.0 115.0 455.0 455.0

**Summary Statements**

Group sample sizes of 246 and 246 achieve 80% power to detect a difference of -115.0 between

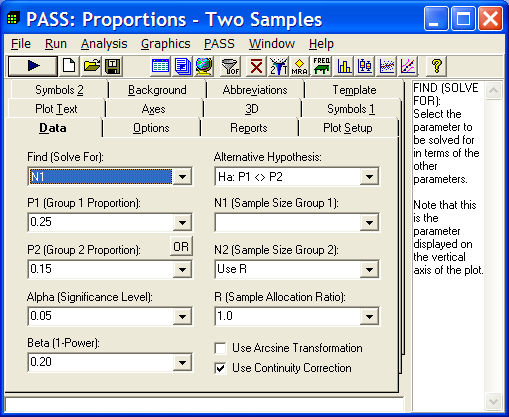
the null hypothesis that both group means are 0.0 and the alternative hypothesis that the mean

of group 2 is 115.0 with known group standard deviations of 455.0 and 455.0 and with a

significance level (alpha) of 0.05000 using a two-sided two-sample t-test.

**b) LBW incidence:**

1. % of babies below normal 25th percentile will decrease with low phytate maize.
2. Eventual analysis: Logistic regression, perhaps adjusting for mother’s weight, gestation time, etc.
3. Simplified analysis: Comparing two proportions.
4. Assume 25% incidence in normal maize group declines to 15% in low phytate group. Do not need SD for proportions, but do need proportion in one group.



**Allocation Odds**

**Power N1 N2 Ratio P1 P2 Ratio Alpha Beta**

0.80003 269 269 1.000 0.25000 0.15000 0.529 0.05000 0.19997

1. Probably slightly too optimistic.
2. Several different values of P1 and P2 were tried.
3. Using two-sided level 0.05 z-tests, to achieve 80% power to detect a decrease from 25% to 15% in LBW incidence between control and low phytate maize groups we require 270 subjects per group. Assuming 15% drop-out, we will enroll 320 mothers per group.

**Example of sample size section for proposal (mean birthweight)**

To estimate sample size for the maternal maize intervention and the infant supplement intervention the following assumptions and estimates were used. Sample size estimates were made using PASS software (PASS User’s Guide, 2000, NCSS: Kaysville, UT). A two-tailed Type I error of 0.05 and a power of 0.80 were used for both trials.

*Maternal maize intervention*

1. Birth records for the period January 1, 2002 to February 10, 2002 at Chimaltenango National Hospital, Guatemala, were obtained by Dr. Manolo Mazariegos. Omitting four birth weights recorded <1000g, this sample yielded 344 live births with mean birth weight 3007g and standard deviation 455g. Two of those omitted were verified to have died at birth, and discussions with clinicians have indicated that birth weights <1000g are not viable, so such values will be excluded in analyses as well.
2. We assume an increase in mean birth weight of 115g to be a reasonable expectation of what can be achieved due to use of low phytic acid maize, and to be of clinical significance to the health of babies and infants.
3. We assume a 15% drop-out rate due to failure to conceive, failure to deliver a live baby, or failure to comply with protocol.

These assumptions and estimates result in a sample size of 300 women enrolled per maize group (total of 600), providing about 250 births per maize group available for analysis (500 total).

**3.6 Common more complex situations:**

* **Growth rates**

**Example:** Compare growth rates (height) months 6-12 between two groups.

**My approach:** Two-sample t-test with outcome = growth rate (cm/month).

**Issues:** How to estimate variability of growth rates from monthly data in previous study?

* **More than two groups**

**Example:** Compare 3 or more groups or ...

**My approach:** Two-sample t-test comparing any two groups.

**Issues:** Consider overall ANOVA? Multiple comparisons?

* **Regression and correlation**

**Example:** Associations among TAZ, FAZ, IZ, EFZ, …

**My approach:** Detectable effect is a correlation between two variables, perhaps with other variables in the model

**Issues:** How many other variables? I don’t find these results terribly enlightening.

* **Factorial designs**

**Example:** Two factors: HF/LF diet and Lean/Overweight

**My approach:** Use statistical theory to calculate variances of comparisons of interest. Trick PASS into doing the calculations.

**Issues:** Which comparisons? HF vs LF for Lean? HF vs LF overall? HF vs LF for lean versus HF vs LF for overweight? Repeated measures on neither or one or both factors?

* **Multiple measurements**

**Example:** Measure each subject for \_\_\_\_ days on each treatment. How many subjects and how many days?

**My approach:** Use more statistical theory to calculate variances of comparisons of interest. Trick PASS into doing the calculations.

**Issues:** Need several variances. Trade-off between more days and more subjects.

* **Linear or unusual trends**

**Example:** Hypothesize a non-linear relation between dietary fat and energy intake. Use 6 levels of dietary fat in an incomplete cross-over design.

**My approach:** Use simulation with several types of non-linear trends.

**Issues:** Which non-linear trend? How many days on each fat level?



* **Diagnostic tests**
* **Multinomial outcome**

**Some questions to ask – Study design, power and sample size estimation**

1. Briefly, describe the study purpose and questions.
2. Design of study, especially are there repeated or paired measurements?
3. What are the primary outcomes on which to base power? (To self: what type of variable are they (numerical, binary)?)
4. Roughly how many subjects will/can you use (a range)?
5. If known, size of detectable difference
6. When do you need it by?
7. Most power and sample size questions require estimates of some quantities:
8. Numerical outcomes require a SD.
9. Binary outcomes require a proportion.

## These can usually be obtained from

1. Published papers
2. Previous (pilot?) data
3. Investigator estimates of ranges (SD ≈ Range / 4)

Published studies or pilot data should satisfy at least:

1. Same outcomes
2. Same species
3. Roughly same timing (eg: hours for outcomes measured over hours)
4. Same design if possible (paired data for paired designs, etc.)
5. I believe they need not have the same treatments, drugs, etc.

Exercise:

3. A study is being designed to test the effectiveness of a new weight loss drug. Subjects will be weighed before the drug is administered and again three weeks afterward. The researcher wants the study to be sensitive enough to detect a weight loss of about 1 kg, and needs to know if the study is viable with 25 subjects. The only paper the investigators could find describing a similar study using a different weight loss program but also over three weeks writes “The 14 subjects lost an average of 1.8 kg (p=0.010)”. Give an appropriate power estimate to help decide if the proposed study is viable.