Error, Power And Sample Size

References:

1. “Designing Clinical Research”. SB Hulley and SR Cummings. Williams & Wilkins, 1988(?).

2. “Sample Size Tables for Clinical Studies”. D Machin, M Campbell, P Fayers and A Pinol. Blackwell Science, 1997.

Special Case: Descriptive Study

- Since there is no hypothesis to test, there is no issue of power.
- However, adequate sample size is still a relevant concern.

Goal: Make the sample size large enough so that the descriptive sample estimates are close to the truth for the population.

Practical Implementation: Make the sample size large enough so that the confidence intervals are sufficiently narrow.

Example: Sensitivity and Specificity

- We propose to study a new diagnostic technique.

- The new technique is good if its sensitivity and specificity are high.

- We choose samples sizes (for the number of people with disease, and the number of people without disease), so that the confidence bounds around the sensitivity and specificity are narrow.

- Since sensitivity and specificity are simply rates (proportions), we need the formula for confidence bounds for proportions:

$$95\%\, CI \left[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \;,\; \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Now, choose n to make the bounds sufficiently tight (you set the standard for how narrow they need to be)

Problem: Before we do the study, we do not know $\hat{p}$ so ... ?

(a) Guess at $\hat{p}$

(b) Use the worst case

$\hat{p} = 1/2$ → most variable and widest interval

$$\left[\hat{p} \pm 1.96 \sqrt{\frac{(1/2)(1/2)}{n}} \right]$$

$$\left[\hat{p} \pm \frac{1.96}{2} \sqrt{\frac{1}{n}} \right]$$

$$\approx \pm \sqrt{\frac{1}{n}}$$

Some Calculations:

Sample Size		Worst Case (p=.5)	Typical Case (p=.9)
n = 5	→	± .45	± .26
n = 10	→	± .32	± .19
n = 20	→	± .22	± .13
n = 100	→	± .10	± .06
n = 200	→	± .07	± .04

Usual Case: Hypothesis Testing

- Your study has the ultimate goal of making a comparison using a statistical test.

- If the statistical test is significant, you will conclude that there is an effect; if the test is not significant, you will conclude that the effect is absent.

Problem: In drawing conclusions (i.e., in making decisions), you could be in error

Goal: Design the study to keep your error rates minimal

Practical Implementation: Carry out some sample size/power calculations during the design of the study.

Types of Calculations:

Pre-Hoc Calculations

1. Determine the sample size necessary to detect a clinically important effect with pre-specified surety.

2. Determine the chance that, with a given sample size, a clinically important effect can be detected.

3. Determine the magnitude of the effect that can be detected, with a pre-specified surety and sample size.

Note: The same formula is used for all three of these purposes.

Post-Hoc Calculations

1. Once the study results are known, determine the chance that the conclusion is correct.
    a. If the study is significant → 5% (usually)
    b. If the study is not significant → ?

2. Provide confidence intervals to interpret non-significant results. (Goodman and Berlin; Annals of Internal Medicine; 1994; Vol 121; pg 200)

Types of Errors

(1) **Type I** = **α-Error** (False Positive)

= Reject the null hypothesis when it is true

(2) **Type II** = **β-Error** (False Negative)

= Fail to reject the null hypothesis when it is false

where

α = $P(\text{Reject } H_0 \mid H_0 \text{ is true})$

β = $P(\text{Do not reject } H_0 \mid H_0 \text{ is false})$

and

Power = $1 - \beta$

= $P(\text{Reject } H_0 \mid H_0 \text{ is false})$

Usual Process

- For many statistical comparisons (i.e., comparing means; comparing rates; comparing censored survival times), there is an equation which involves α, power, sample size, and the effect size.

- You have to specify three of these quantities.

- You then use the formula to solve for the fourth.

Convention: We first choose α, the false positive rate, to be as large as we can tolerate. (Almost always, α=5%)

Implication: If you have a positive finding (i.e., you reject the null hypothesis) there is an α% chance that you are wrong.

Reason: False positive results are usually considered the most critical sorts of errors since they may change clinical practice for no good reason.

Then: We specify 2 of the following 3 quantities and use the formula to calculate the third:

1) Sample size 2) Desired power 3) Effect size

Some Formulas For Calculating Sample Sizes
Two-Group Comparisons, With Equal Sample Sizes In Each Group

1. Comparing Two Proportions: n is the required sample size per arm

$$n = \frac{[\, z_{1-\alpha/2}\sqrt{2\bar{p}(1-\bar{p})} + z_{1-\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}\,]^2}{(p_2-p_1)^2}$$

Where $\bar{p} = (p_1 + p_2)/2$

Note 1: While the sample size requirement is most sensitive to the rate difference, it is also sensitive to the individual values of p_1 and p_2.

Note 2: It is not sufficient to specify just the relative risk or the odds ratio.

Some Formulas For Calculating Sample Sizes
Two-Group Comparisons, With Equal Sample Sizes In Each Group

2. Comparing Two Means: n is the required sample size per arm

$$n = \frac{2 s_p^2 (t_{n+m-2,1-\beta} + t_{n+m-2,1-\alpha/2})^2}{(\mu_1 - \mu_2)^2}$$

$$\approx \frac{2 s_p^2 (z_{1-\beta} + z_{1-\alpha/2})^2}{(\mu_1 - \mu_2)^2}$$

Note 1: To be perfectly, technically correct, the sample sizes should be calculated using constants from the t-distribution, but this requires knowing n, which makes the process circular. Therefore, we initially use constants from the Normal table.

Note 2: If the sample size requirement (based on the Normal table) is small (say less than 20), then you may want to repeat the calculation, this time using constants from the t-table, with degrees of freedom based on the initial sample size calculation. The sample size requirement will go up after the second calculation.

Note 3: The formula requires s^2. Ideally you would get this from pilot data. Alternatively, you could estimate s by 1/4 of the usual data range.

How Do You Specify The "Desired Power"?

1. Usually, set the power to be 90%

   - This is the usual standard for most confirmatory studies, especially those competing for NIH funding.

2. If your study is meant to provide the final and conclusive word in investigating a therapy where prior large trials have given conflicting results, set the power to be 95%

3. Set the power at 80% or 85% if your study is a pilot, or exploratory, or more epidemiological in nature (i.e., you want to find all dietary risk factors for heart disease).

4. Do not set the power below 80%.

Statistical Implication: When you set the power, you are specifying a value for a constant in the power/sample size formulas that follow. Those constants are:

Power=95% → $z_{1-\beta} = 1.65$
Power=90% → $z_{1-\beta} = 1.28$
Power=85% → $z_{1-\beta} = 1.04$
Power=80% → $z_{1-\beta} = 0.84$

Notation: $z_{1-\beta}$ represents the point on the x-axis of the Normal distribution, chosen so that (1-β)% of the area of the Normal curve is to the right of $z_{1-\beta}$

Example: $z_{.975} = 1.96$ (here, β is 2.5%)

How Do You Specify The "Effect Size"?

Definition: The definition of an "effect size" depends on the kind of data we will be analyzing and the type of statistical test that will eventually be used. As examples, the effect size could be: the difference between two means; the difference between two proportions; the ratio of mean survival times when there will be censoring.

Specification:

1. Ideally, from pilot data that you collect, or that is reported in the literature.

2. Based on your best guess.

3. Based on the principle of "minimum clinical significance".
   - i.e., what is the smallest effect of your intervention that would be clinically meaningful (therefore worth doing a study to detect)?

4. Resort to more “generic” measures of effect.

   a. Standardized Effect – for comparing two means
      where, Standardized Effect = $(\text{mean}_1 - \text{mean}_2) / \text{sd}$

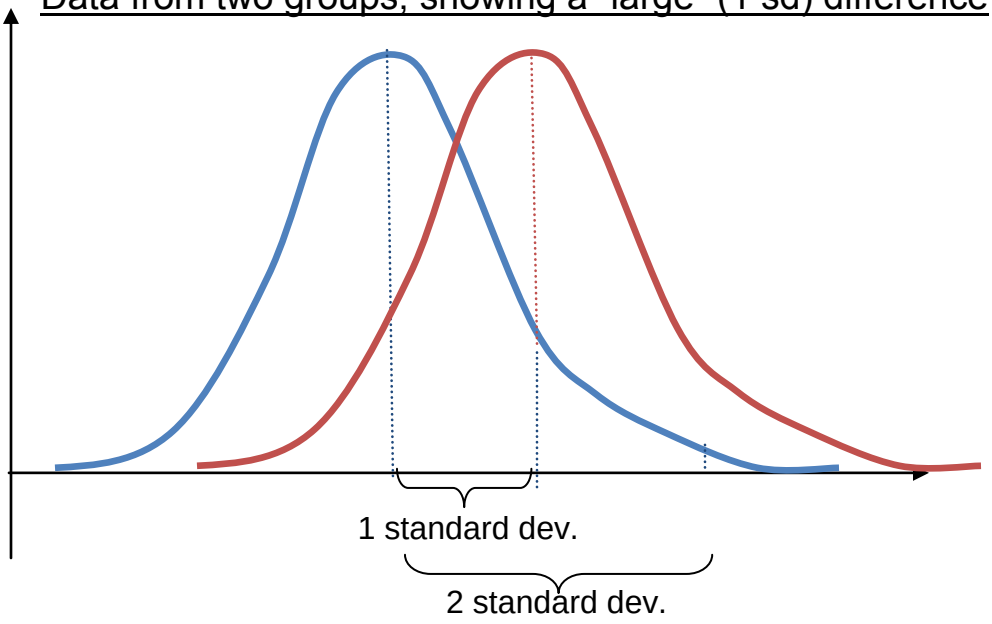
      Advantage: This reduces the sample size formula to:
      $$n = 2 (z_\alpha + z_\beta)^2 / (\text{Standardized Effect})^2$$
      which requires no knowledge about the means or the sd

"Accepted" Standards:

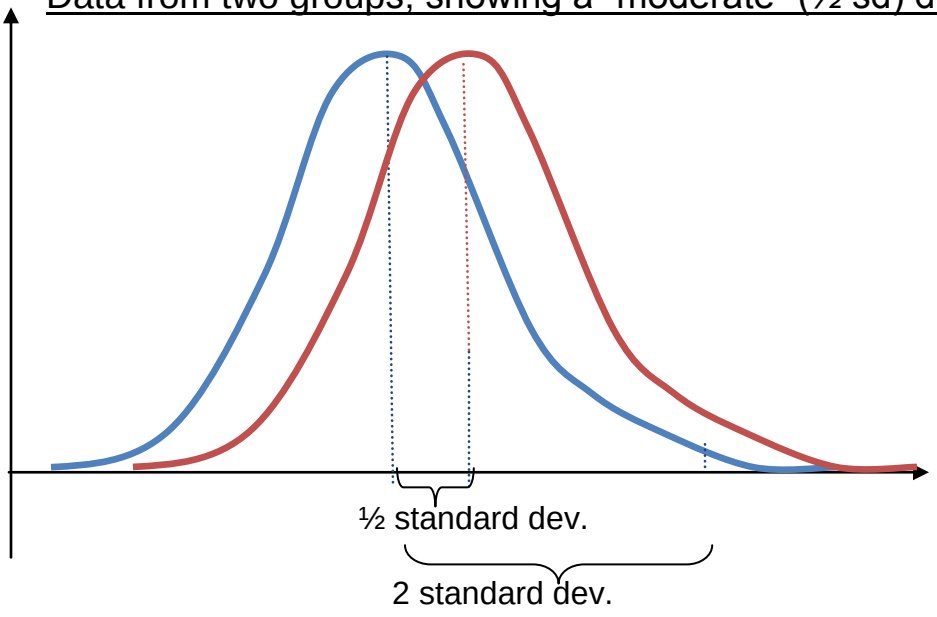
Standardized Effect = ¼ → Small Effect
Standardized Effect = ½ → Moderate Effect
Standardized Effect = 1 → Large Effect

Usual Language: With 84 patients per group, it will be possible to detect moderate effects of intervention with 90% power.

Data from two groups, showing a "large" (1 sd) difference



Data from two groups, showing a "moderate" (½ sd) difference



Calculations for Standardized Effects

Standardized Effect Size	Sample Size Required Per Arm For α=.05 and Power: 80%	90%
.10	1570	2102
.15	698	934
.20	393	526
.25	251	336
.30	174	234
.40	98	131
.50	63	84
.60	44	58
.70	32	43
.80	25	33
.90	19	26
1.00	16	21

b. Correlation Coefficient -- for association between 2 continuous measures

Advantage: This reduces the sample size formula to a statement about the strength of the correlation and requires no knowledge about the means or the sd's

"Accepted" Standards:
Correlation Coefficient $\leq$ 0.4 → Small Effect

Correlation Coefficient 0.4 to 0.7 → Moderate Effect

Correlation Coefficient $\geq$ 0.7 → Large Effect

Usual Language: With 62 patients, it will be possible to detect a moderate relationship between the predictor and the outcome with 90% power.

Calculations for Correlation Coefficients

	Sample Size Required For α=.05 and Power:	
	80%	90%
Correlation		
.05	3134	4200
.10	782	1047
.15	346	463
.20	194	259
.25	123	164
.30	85	113
.35	62	82
.40	47	62
.45	36	48
.50	29	38
.60	19	25
.70	13	17
.80	9	12

Typical Calculation 1: Power for Case-Control Studies

Assume: H_a: $p_1 = .06$ and $p_2 = .16$

# Cases	# Controls		
100	versus 100	→	62% Power
100	versus 200	→	77% Power
150	versus 150	→	79% Power
100	versus 300	→	83% Power
100	versus 400	→	85% Power
100	versus 500	→	87% Power
100	versus ∞	→	93% power

Typical Calculation 2: Sample Size for Case-Control Studies

Assume: Control arm mortality = 20%

Treatment arm mortality = 10%

Power = 90%

Type I Error = 5% (2-sided)

Case/Control Ratio:	Sample Size Required # Cases	# Controls	Total
1:1	286	286	572
1:2	210	420	630
1:3	184	552	736
1:4	171	684	855
1:5	163	815	978

Typical Calculation 3: Sample Sizes for Comparing Event Rates

Assume: Control arm mortality = 40%

Type I Error = 5% (2-sided)

Treatment Arm Mortality:	Sample Size Required Per Arm To Achieve The Following Power: 95%	90%	80%
38%	15,554	12,596	9,435
35%	2,473	2,008	1,510
30%	608	496	376
25%	264	216	165
20%	143	118	91
10%	58	48	38

Typical Calculation 4:

Detectable Effects When Comparing Event Rates

Assume: Control arm remission rate = 10%

There will be the same number of treated and control patients

Type I Error = 5% (2-sided)

Table: Required Remission Rate in Treated Patients
(Assuming a 10% Remission Rate in Control Patients)

	In Order to Achieve the Following Power:		
	80%	85%	90%
Patients Per Arm:			
100	>26%	>27%	>29%
150	>23%	>24%	>25%
200	>21%	>21%	>22%
250	>19%	>20%	>21%
300	>18%	>19%	>20%

Special Topics In Sample Size and Power

1. Drop-Outs

This refers to the problem of patients who enroll in the study, but who do not provide any analyzable data because they drop out of the study. (This does not include censored data.)

Solution: Use the usual sample size formulas to calculate the number of subjects required. Then increase this number by the appropriate percentage to insure that a sufficient number of patients will provide usable data.

Example: The sample size formulas show that you need 100 patients in the study. However you expect a 20% drop-out rate. Therefore, enroll 125 patients. At the end of the study, you will still have usable data for 100 subjects, as required.

Special Topics In Sample Size and Power

2. Adjustment for Confounders / Multiple Regression Analysis

You are doing a cohort study, not a randomized trial, and you suspect that there will be confounders. The primary analysis will involve multiple regression, adjusting for these confounders.

Non-Solution: There is no good approach to sample size calculation for regression. These would require a great deal of a priori knowledge as to the impact of each predictor on the outcome, and the inter-relationships between the predictors. If these relationships could be quantified, then computer simulation is typically used to estimate power.

Available Solution: Common packages will calculate the power (or sample size) associated with a given increase in the r-squared of a multiple regression model.

For example, with 100 subjects, if you have 3 covariates in your model with an r-squared of 30% and if the predictor of interest increases the r-squared by 5%, then you will have 78% power to detect this effect of the predictor.

➔ Is this an intuitive/useful calculation?

Common Solution: Do the simpler group-versus-group calculations that we have reviewed. However, when you plug in the effect estimate, use a number that is more conservative (smaller), reflecting the impact of your predictor/risk factor/treatment **after** adjustment for confounding.

Example: Preliminary data suggest that the crude, unadjusted effect of a risk factor will be to increase disease incidence from 10% to 20%. However, people with this risk factor also engage in many other high-risk behaviors which could account for part of this increased incidence of disease. Therefore, when you do your sample size calculations, make a more conservative estimate that the risk factor increases disease incidence from 10% to 16%.

Special Topics In Sample Size and Power

3. Three-Group Comparisons / Multi-Group Comparisons

Your primary analysis is a comparison of 3 or more groups (i.e., ANOVA). However, all of the formulas in class only deal with 2-group comparisons.

One Solution: Specify the three means and proceed with the power/sample size calculation for the ANOVA. The proper formulas can be found in:
J. Cohen; "Statistical Power Analysis for the Beharioral Sciences"; 1988.
or J.L. Fleiss; "The Design and Analysis of Clinical Experiments"; 1986.

Usual Solution: This assumes that after you find significance on the ANOVA, you will still be interested in comparing between individual groups to find out where the actual differences lie (i.e., in "pairwise testing"). Since each pairwise test is simply a 2-group comparison, use the formulas already available from class. Otherwise, you may end up with enough power for the ANOVA, but not for the subsequent pairwise comparisons. You may/should adjust the α-level to reflect the
number of pairwise comparisons you intend to carry out: in 3-group ANOVA, there can be as many as 3 pairwise comparisons, so use an α of .05/3=.0167

Special Topics In Sample Size and Power

4. Non-Compliance and Cross-Overs

You have an estimate of the effect of your intervention in a perfect world where everyone assigned to therapy receives 100% of the intended dose, and everyone in the control arm receives placebo. However, in the real world, some of the intervention patients are non-compliant, and some of the placebo patients seek out alternative therapies.

Solution: Make some assumptions as to the extent of non-compliance and cross-overs, and adjust your effect estimate toward the null hypothesis to reflect the real world.

Example: In a perfect world, intervention patients will improve by 10 points on quality of life and control patients will improve by only 2 points (placebo effect). However, you expect 20% of the intervention patients to terminate treatment early, thereby having their improvement reflect only the placebo effect. In addition, 30% of the control patients will seek out alternative therapies which will improve their quality of life by a moderate 5 points.

Overall, the intervention patients will show an improvement of:

(80%x10 points) + (20%x2 points) = 8.4 points

while the control patients will show an improvement of:

(70%x2 points) + (30%x5 points) = 2.9 points

Special Topics In Sample Size and Power

5. Multiple Endpoints / Multiple Predictors

You intend to run many analyses, looking at multiple outcome measures and multiple predictors.

Solution: For each outcome and each predictor, run the appropriate sample size calculation. Do this for every outcome/predictor pair. The necessary sample size for the study is the maximum sample size across these calculations. (Note, it is not the sum of the sample sizes, only the maximum.)

You may/should adjust each calculation so that the alpha-level is divided by the number of comparisons intended in total.

If you find that one outcome/predictor comparison requires a very large sample size, you may elect to take this comparison out of your primary hypotheses and thereby allow for the study to proceed with a smaller sample size requirement.