

Power analysis and sample size calculations for clinical trials

Sandra Taylor, Ph.D.
February 8 & 15, 2017

UCDAVIS
CLINICAL AND TRANSLATIONAL
SCIENCE CENTER

UCDAVIS
MIND INSTITUTE

UCDAVIS
COMPREHENSIVE
CANCER CENTER

UCDAVIS
ENVIRONMENTAL HEALTH
SCIENCES CENTER

**We are video recording this
seminar so please hold
questions until the end.**

Thanks



How many subjects do I need?



This is one of the most common questions we get. “How many subjects do I need?” My intent today is to provide a very basic and shall we say “gentle” introduction to power and sample size issues.

Provide understanding of

- **Why** sample size calculations are important
- **Where** sample size calculations fit into study planning
- **What** information is necessary to estimate sample size needs
- **How** to conduct simple sample size calculations
- **When** to seek help from a statistician

More specifically, I want to provide an understanding of why sample size calculations are important, where they fit into study planning, what information I necessary to estimate sample sizes, how to conduct simple sample size calculations, and when to seek help from a statistician.

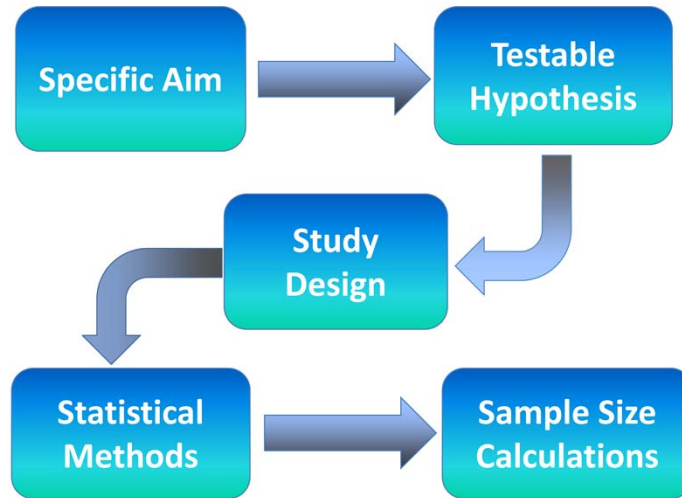
Why are sample size calculations important?

- **Enroll too few – insufficient power to detect differences**
- **Enroll too many – unnecessary costs**

Recognizing the importance of sample size and power assessments in study development, the UCD IRB now prompts investigators to provide a power analysis in its Protocol Template.

So, why are sample size calculations important? I suspect the reasons are evident but in case they are not, determining sample size needs BEFORE initiating the study is important because if you enroll too few of subjects, then you will not have sufficient statistical power to detect differences you are interested in. In essence you will have wasted your time and resources. On the other hand, if you enroll more subjects than you need, you will incur unnecessary costs and time and in the case of a human clinical trial you could end up delaying adoption of the most effective treatment. The UCD IRB recognized the importance of sample size and power assessments in study development and now prompts investigators to provide a power analysis in its Protocol Template. This was actually the impetus for this talk as we started to get more folks coming in asking for help in addressing this.

Where do sample size calculations fit in the research process?



Oftentimes, investigators come to us when they have their study all planned and say “Ok, now how many subjects do I need” or perhaps the IRB kicked back their application asking for a justification of why 6 is enough. I have to say that I am still surprised when investigators come in and say that they are using a certain number because that is what we have always used. Sample size calculations depend on a series of steps/decisions made as you develop your study. It starts with the specific aim for your study. This then needs to be translated into a testable hypothesis. What specifically are you going to measure and compare or evaluate? These then feed into the study design. Together the study design and the specific hypotheses to be tested determine the statistical methods that will be used which then form the basis for the sample size calculations. Thus, sample size calculations are intricately linked with the specifics of the study.

How is sample size determined?

Depends on:

- **Specific aim – primary hypothesis of the study**
- **Study design**
 - These two influence the statistical test.
- **Effect size to be detected**
- **Variability of the response variable**
 - Researchers need to provide this information

There is no magic number.

So, how do we determine sample size requirements? As shown in the previous slide, the specific aim of the study is critical. This really comes down to what is the specific hypothesis that you want to test. And then secondly, the study design comes into play. In addition to those kind of study design aspects, we also need to know the effect size to be detected and the variability of the response variable. The investigators need to provide this information. I will get into this aspect in a more detail because it seems to be an area of misunderstanding. Not commonly, investigators show up and ask how many do I need and seem quite unprepared when we start asking them to provide information on effect size and variability. It's like some folks think there is some magic number when in fact there isn't .

What information is needed?

- **Desired power – Prob. rejecting H_0 if it is in fact false.**
 - Typically 80 – 90%
- **Allowable Type I error – Prob. rejecting H_0 when it is true.**
 - Typically 5%
- **Effect size to be detected**
- **Variability of the response variable**

Let's say that we have a well defined testable hypothesis and a good study design and we are ready to do some sample size calculations. What information is needed to do this? There 4 specific bits of information that are needed. First, what power do you want. Power is the probability of rejecting the null hypothesis when in fact it is false. You want this to be high and we typically set that at 80 or 90%. Second, what level of type I error are you willing to accept? Type I error is the probability of rejecting the null hypothesis when in fact it is true. This is typically taken to be 5%. Third and fourth we need effect size and variability of response.

How do these affect sample size requirements?

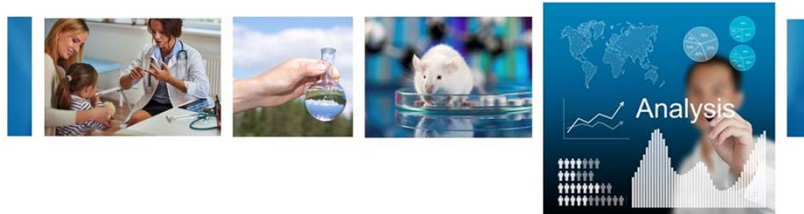
- **Power: Higher** → **Larger N**
- **Type I Error: Higher** → **Smaller N**
- **Variance: Larger** → **Larger N**
- **Effect Size: Larger** → **Smaller N**

How do each of these affect your sample size calculations? Well, the higher you want your power to be, the larger your sample size needs to be. Conversely, the more Type I error you are willing to accept, the smaller your sample needs to be. The more variable your response (ie., larger variance) the larger your sample size will need to be and finally, the larger the effect size you are interested in being able to detect the smaller sample size you will need. This is an important one to think about from the other angle. If you want to be able to detect a small effect size, you will need a large sample size.

Where do I get effect size and variance information?

- **Published results on same outcome**
 - Often source for variance and baseline levels
- **Published results for similar outcome**
- **Pilot study data**
- **Expected treatment effect**
- **Clinically meaningful change**

When you say I need to know the variance of my response variable and the effect size I want to detect, where do I find that information? If I knew this information I wouldn't need to do the study. There are a number of possible sources. First, published for the same outcome can give information on the variance and often baseline levels. If there is nothing on your specific outcome you might be able to find information on a similar outcome. You might be fortunate to have pilot study data which could give you information on both the variance and the effect size. For effect size, you can think about what you expect or hope to see. Maybe you are investigating a new medication and you know the average effect of the current medication. The new medication might only be of interest if it performs at least as well as the current medication and so you could use the average effect size of the current medication. Finally, in cases where there just isn't any information to go on for effect size, we simply ask the investigator, what would be a clinically meaningful change? At what point would you say, this is something of interest?



Yes, but **how** do you actually do a sample size calculation?

UCDAVIS
CLINICAL AND TRANSLATIONAL
SCIENCE CENTER

UCDAVIS
MIND INSTITUTE

UCDAVIS
COMPREHENSIVE
CANCER CENTER

UCDAVIS
ENVIRONMENTAL HEALTH
SCIENCES CENTER

By this point you may be feeling that this all rather conceptual and thinking how do I actually do a sample size calculation? So, now I am going to run through some specific examples.

Determine method to use

- **What type of variable is the outcome?**
 - Continuous, categorical, survival
- **How many groups do you have?**
 - One, two, > two, continuous predictors

OUTCOME	1 GROUP	2 GROUPS	> 2 GROUPS
Continuous	One-sample t-test	Two-sample t-test	ANOVA
Categorical	One-sample Proportion test	Two-sample Proportion Test	Logistic Regression Chi-square Test
Survival	Kaplan-Meier Estimate	Log Rank Test	Log Rank Test

First off, you have to determine what method to use. This amounts to what statistical test is appropriate for your outcome type and what your specific hypothesis is. First question is what type of variable is the outcome? Is it continuous, categorical or survival data? How many groups do you have or do you have a regression problem? This chart can help you determine the appropriate method for studies with groups. Regression problems are trickier and I am keeping this talk to simple situations. So, what we see is that if we have a continuous outcome and two groups we can use a two-sample t-test. More than two groups and we need to consider an ANOVA. Similarly there are different tests for categorical outcomes and survival outcomes.

Consider a Weight Loss Study

- **Two groups**
 - 1) Self-guided Program, 2) Coached Program
- **Outcome of interest**
 - Amount of weight lost (kg)
- **Continuous outcome with two groups**
 - Two-sample t-test
- **What effect size to detect?**
- **What is the variance?**

Let's see how we would go about a sample size calculation. Consider weight loss study. Say you are interested in whether people do better with a self-guided program or with program that includes, regular one-one coaching from a real live person. The outcome of interest is the amount of weight lost in kg between the start and end of the program. So, we have a continuous outcome with two groups for which the previous chart tells us to use a two-sample t-test. Now we need to specify the effect size and the variance.

Pelligrini et al. 2011. Comparison of a technology-based system and an in-person behavioral weight loss intervention.

- **6-month weight loss**

- TECH program, SBWL only program and TECH+SBWL program

“Weight loss among those who completed both baseline and 6-month assessments was significant for SBWL (-7.1 ± 6.2 kg), SBWL+TECH (-8.8 ± 5.0 kg), and TECH (-7.6 ± 6.6 kg).”

- **Standard deviations for change in weight were 6.2, 5.0, 6.6**

A published study by Pelligrini looked at weight loss with similar programs. This group had a technology only group, a technology and in-person consultation group and a in-person consultation only group. They found the following mean weight changes and standard deviations. This is exactly what we need.

Assumptions for Calculations

- **Standard Deviation of the Change**
 - 6.0 kg
- **Effect Size to Detect**
 - How about 4 kg?
- **Power**
 - 90%
- **Type I error**
 - 5%

Based on this information, what should we use in our calculations. For the standard deviation of the change, how about 6 kg. What effect size do we want to be able to detect? They found mean changes of 7 to 9 but perhaps we would like to be able to detect a smaller change, say 4 kg. Let's take power to be 90% and the Type I error to be 5%. This is all the information that we need. Now we can go to an on-line calculator to get the sample size estimate. There are many available. I am going to show you the Southwest Oncology Group's website. You can get to this website from the Biostat section of the CTSC website.

Southwest Oncology Group Statistical Tools Website

Two Arm Normal

Two Arm Normal is a program to calculate either estimates of sample size or power for differences in means. The program allows for unequal sample size allocation between the two groups.

User Input | **Program Output**

Select Calculation and Test Type

Sample Size | 1 Sided
 Power | 2 Sided

Select Hypothesis Test Parameters

Mean: Arm 1	Mean: Arm 2	Standard Deviation	Ratio of Sample Size Arm2/Arm1	Alpha
0	4	6	1	0.05

Power | **Total Sample Size**

0.9	95
-----	----



Calculate

[Help Document](#)

<https://stattools.crab.org/>

So, this says that we would need a total sample size of 95 (i.e., 48 in each group) to detect a 4 kg change with 90% power given the other assumptions.

Sample size calculations are not absolute.

- **What if 48 subjects per group was too expensive?**
 - 20/group  Power = 56%
 - Too low to detect 4 kg difference.
- **What difference can be detected at 80% power with n=40?**
 - 5.5 kg difference  Power = 83%
 - Are these values acceptable?

Based on a two-sample t-test, 20 subjects per group will provide 83% power to detect a 5.5 kg difference in weight loss between the two groups assuming a standard deviation of 6.0 kg (Pelligrini et al. 2011).

What if you couldn't afford to have 48 subjects in each group? One concept I would like to get across is that sample size calculations are not absolute. If 48 subjects was not at all doable we should take a look at the assumptions we made in doing the calculations. Suppose we could only afford 20 subjects per group. We can use SWOG Stat to calculate what power we would have with this. We would only have 56% power to detect a 4 kg difference which is too low. We can also look at what difference can be detected for n=40 and power of 80% by doing some "gaming". Doing this we find that with n=40 we have 80% power to detect a 5.5 kg difference. You as an investigator then need to determine whether this is acceptable or whether you really need to be able to detect a 4 kg difference. The statistician does not have the answer to this. Assuming this is ok, you could put in your proposal or IRB application, the following statement to justify your sample size.

“Adequate” sample size does not ensure “significant” results

- **Observed effect size < 5.5**
 - Not significant
- **Standard deviation > 6**
 - Not significant
- **Power = 83%**
 - 17% chance of not rejecting when null is false
- **Type I error = 5%**
 - 5% chance of rejecting true null

Let's say that you have diligently went through and did an appropriate sample size calculation and conducted the study according to plan and find nothing significant. Don't shoot the statistician. It is important to recognize the uncertainty inherent in sample size calculations any one of which could result in you not having significance. First, if the effect size is less than assumed (i.e. less than 5.5) you won't detect it. Second, if the standard deviation is actually greater than assumed, you would miss an effect size of 5.5. Third, power was 83%. That means there is a 17% chance of not rejecting when the null is false. That's not a trivial amount. Finally, it is important to remember that with a 5% type I error, there is a 5% probability of rejecting the null even though it is actually true.

Consider an Outreach Program to Increase Follow-up Visits for Abnormal Pap Smears

- **Two groups**
 - 1) Usual care, 2) culturally-sensitive outreach
- **Outcome of interest**
 - Rate of follow-up (proportion)
- **Categorical outcome with two groups**
 - Two-sample proportion test
- **What proportion follows-up under usual care?**
- **What effect size is meaningful or likely?**

I want to quickly go through a calculations when the outcome is a proportion because this is a pretty common outcome. Perhaps the interest is in getting women who come to a county clinic and have an abnormal pap smear to come to a follow up visit. Again, let's consider two groups – one that receives usual care and the other gets culturally sensitive outreach. Our outcome of interest is in the proportion of women who come back for recommended follow up. Since it is a proportion with two groups we will use a two-sample proportion test. To do the sample size calculations, we need to have an estimate of the proportion that currently follow up under the usual care and then we need to say what effect size is meaningful.

Assumptions for Calculations

- **For proportions, only need to know “baseline” proportion and effect size**
- **Current proportion of women with abnormal Pap smears that follow up**
 - Assume known to be around 40%
- **Effect size to detect**
 - Suppose increase to 60% would be clinically meaningful.

For proportions we only need to know the baseline and the effect size. Suppose we know that current proportion that follow is around 40% and that we believe it would be clinically meaningful if we were able to increase that to 60%. Again we can use SWOG Stat to estimate the sample size.

Southwest Oncology Group Statistical Tools Website



STATISTICAL TOOLS DESIGN ANALYSIS PROBABILITIES ABOUT US

Two Arm Binomial

Two Arm Binomial is a program to calculate either estimates of sample size or power for differences in proportions. The program allows for unequal sample size allocation between the two groups.

User Input Program Output

Select Calculation and Test Type

Sample Size 1 Sided
 Power 2 Sided

Select Hypothesis Test Parameters

Control Group Proportion	Experimental Group Proportion	Alpha	Sample Size Ratio 2-to-1
0.4	0.6	0.05	1

Calculate Power/Sample Size

Power	Sample Size
0.9	280

[Help Document](#)

<https://stattools.crab.org/>

SWOG Stat tells use that we need 280 (140 per group) under these assumptions. As we did for the weight loss study, you as the investigator would have to make a determination of whether you could recruit this number of participants or if relaxing some of the assumptions could be justified.

One-sample Studies

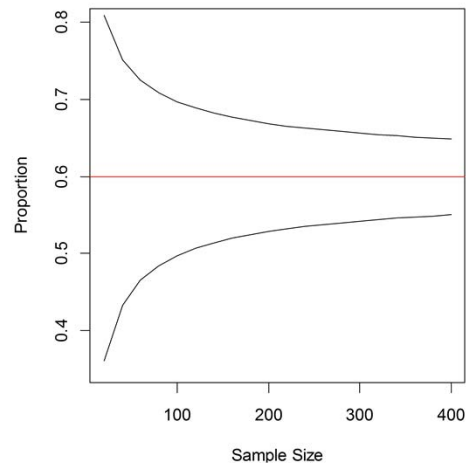
- **Compare hypothesized value**
 - One-sample proportion or t-test
- **Characterize a trait or occurrence in population of interest**
 - Objective for sample size calculation is number of subjects required for desired precision

Lastly, I want to touch on one-sample problems because they tend to create confusion. There are two general cases where you might have just one group rather than 2 or more. First, is if you want to compare your population to a hypothesized, known or standard value. Perhaps it's a performance measure such as proportion of patients getting an infection during a hospital stay and you want to compare your hospital to the national average or a national standard. Then you would use a one-sample proportion test or if your outcome was continuous a one-sample t-test and proceed in the same manner as for the two sample problems. The other situation is where you want to characterize a trait or occurrence in a population. This is an estimation problem (i.e., you want to estimate a parameter in your population). Then the objective for the sample size calculation is to determine the number of subjects needed for a desired level of precision of the estimate. Larger sample leads to narrower confidence limits (i.e., more precise estimate).

Suppose we are interested in estimating the proportion of patients receiving a particular medication who improved.

Assume $p_0 = 0.6$.

How does precision of this estimate vary with n ?



Suppose we are interested in estimating the proportion of patients receiving a particular medication who improved. Let's say we think the true value is around 60%. This plot shows how the confidence limits change as the sample size increases. Two other things I would like to point out for proportions, is first that the size of the sample also influences the granularity with which you can estimate the proportion. For example, if you only have 10 subjects then the possible point estimates can only be in increments of 10. Second, assuming you want to observe the events you will need to have a sufficiently large sample size to be a reasonable chance of observing the event. So if something has a 1% probability of occurring, with a sample size of 10 or even 100 you very likely will not have any events.

Example: CAIRO4 Study

- **Primary research goal:** Determine whether performing surgery of the primary tumor followed by systemic therapy improves survival in a certain patient population, compared with systemic therapy only.
- **Patient population:** Patients with synchronous unresectable metastases of colorectal cancer and few or absent symptoms
- **Primary outcome:** Overall survival
- **Study design:** Multi-center randomized phase III trial.

BMC Cancer 2014; 14:741

Survival/time to event data is different from binary or continuous data because of censoring.

Time-to-event variables

- **The log rank test is often used to compare two survival curves.**
- **Most sample size calculations assume an exponential survival distribution.**
- $S(t) = e^{-\lambda t}$, where
 - t = time,
 - $S(t)$ = probability of survival to time t , and
 - λ = hazard rate = risk of an event per time unit

Converting parameters: exponential survival distribution

- Hazard rate: number of events per unit time
- Median survival time = $\log_e(2)/(\text{hazard rate})$
- Hazard rate = $\log_e(2)/(\text{median survival time})$
- Hazard rate = $-\log_e(S(t))/t$,

where $S(t)$ = probability of surviving to time t
= expected proportion without an event by t

Sometimes you need to convert among different parameters

Sample size formula

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 [\phi(\lambda_C) + \phi(\lambda_I)]}{(\lambda_I - \lambda_C)^2}$$

where $\phi(\lambda) = \frac{\lambda^2}{1 - [e^{-\lambda(T-T_0)} - e^{-\lambda T}] / \lambda T_0}$

n = number per group

λ_I = hazard rate in intervention group

λ_C = hazard rate in control group

T = total time of trial (first entry to end of study)

T_0 = recruitment time (first entry to last entry)

Example: CAIRO4


- **Null hypothesis**

- Overall survival is not affected by surgery of the primary tumor before systemic therapy in this patient population.

- **Alternative hypothesis**

- Surgery of the primary tumor improves overall survival in this patient population.
- 

CAIRO4: Assumptions for sample size calculations

- **Level of the test: 0.05 (2-sided)**
 - **Power: 80%**
 - **Median survival in control group: 13 months**
 - **Median survival intervention group: 19 months**
 - Minimal difference to justify a surgical procedure
 - **Recruitment period: 36 months**
 - **Minimum follow-up: 6 months**
- 

SWOG Stat Example



STATISTICAL TOOLS DESIGN ANALYSIS PROBABILITIES ABOUT US

Two Arm Survival

Two Arm Survival is a program to calculate either estimates accrual or power for differences in survival times between two groups. The program allows for unequal sample size allocation between the two groups. The survival time estimates also allow for multiple strata or risk groups.

User Input	Program Output
------------	----------------

Select Parameters

Type calculation <input checked="" type="radio"/> Sample Size <input type="radio"/> Power	Type input <input type="radio"/> Hazard Rates <input checked="" type="radio"/> Survival Proportion	Sided <input type="radio"/> 1 Sided <input checked="" type="radio"/> 2 Sided
---	--	--

Number strata 1	Proportion in standard group 0.5	Alpha 0.05
--------------------	-------------------------------------	---------------

36 months = 3 years

Years of accrual 3	Years of follow-up 0.5	Accrual rate 120.01	Hazard ratio 1.46	Total accrual 360	Power 0.8
-----------------------	---------------------------	------------------------	----------------------	----------------------	--------------

Calculate

6 months = 0.5 years

You need to calculate this

50% = median survival

Median survival = 13/12

Stratum	Proportion	Hazard rate, std.	Hazard rate, exp.	Proportion surviving	Survival time
1	1	0.64	0.438	0.5	1.083

Calculating Hazard Ratio

Control Group

- Median survival control group is 13 months = 1.083 years
- Hazard rate = $\log_e(2)/(\text{median survival time}) = 0.64$

Intervention Group

- Want to detect 6 month difference = 19 months = 1.58 years
- Hazard rate = 0.438

Hazard Ratio = $0.64/0.438 = 1.46$



SWOG Stat Example



STATISTICAL TOOLS DESIGN ANALYSIS PROBABILITIES ABOUT US

Two Arm Survival

Two Arm Survival is a program to calculate either estimates accrual or power for differences in survival times between two groups. The program allows for unequal sample size allocation between the two groups. The survival time estimates also allow for multiple strata or risk groups.

User Input		Program Output			
Select Parameters					
Type calculation <input checked="" type="radio"/> Sample Size <input type="radio"/> Power		Type input <input type="radio"/> Hazard Rates <input checked="" type="radio"/> Survival Proportion		Sided <input type="radio"/> 1 Sided <input checked="" type="radio"/> 2 Sided	
Number strata 1		Proportion in standard group 0.5		Alpha 0.05	
Years of accrual 3	Years of follow-up 0.5	Accrual rate 120.01	Hazard ratio 1.46	Total accrual 360	Power 0.8
Calculate					
Hazard Ratio					
Stratum	Proportion	Hazard rate, std.	Hazard rate, exp.	Proportion surviving	Survival time
1	1	0.64	0.438	0.5	1.083

Sample size calculations may not be straight-forward

- **More complex designs require more complex calculations**
- **Examples:**
 - Longitudinal studies
 - Cross-over studies
 - Correlation of outcomes
- **Sometimes simulations are required**
- **Recognize when your situation is not straight-forward**

When in doubt – ask.

You may now have the feeling that there's nothing to doing sample size calculations. This is easy. If that's your feeling – great. That's what I was going for. My hope is that you will feel empowered to do these calculations on your own. However, and this is a big however, not all studies are that straight-forward. Sometimes, sample size calculations can be quite complex and involved. More complex designs necessitate more complex calculations. These design include longitudinal studies, cross-over studies, and studies where there is some kinds of correlation among the outcomes – for example, patients at the same hospital in a multi-center study. Sometimes there isn't pre-packaged software and we need to do some simulations. What is most important is for you to recognize when it is not a straight-forward calculation and to seek help in a timely manner from a statistician. When in doubt – ask.

Help is Available

- **CTSC Biostatistics Office Hours**
 - Every Tuesday from 12 – 1:30 in Sacramento
 - Sign-up through the CTSC Biostatistics Website
 - **EHS Biostatistics Office Hours**
 - Every Monday from 2-4 in Davis
 - **Request Biostatistics Consultations**
 - CTSC - www.ucdmc.ucdavis.edu/ctsc/
 - MIND IDDRC - www.ucdmc.ucdavis.edu/mindinstitute/centers/iddrc/cores/bbrd.html
 - Cancer Center and EHS Center
- 