PharmaSUG Paper ST176

Practical Perspective in Sample Size Determination

William Coar, Axio Research

ABSTRACT

Experiments advance science. They are designed to answer specific questions. A well designed experiment accompanied by appropriate statistical methodology should yield scientifically sound results to answer specific questions with a reasonable amount of certainty. To achieve this, the concepts of type 1 error, power, and sample size are introduced into the experimental design.

While these abstract concepts are based solely on assumptions, they are critical to the integrity of study results. SAS® provides numerous procedures to assist with these parts of experimental design. The designs in clinical research range in complexity, as do the SAS procedures that support them.

This presentation will approach sample size determination from a practical perspective. Even if a set of assumptions are reasonable, they may not result in a feasible sample size. We use a controlled clinical study to demonstrate that sample size and study design can evolve due to practical considerations. The endpoint is 28-day survival for an often fatal medical condition. It is assumed that 50% of patients will survive within 28-days under the standard of care. However, a medical procedure could possibly increase this to 80%. This presentation will discuss the use of PROC POWER, PROC SEQDESIGN, and even simulation to assist with the study design, and how practical considerations may cause the design to evolve.

INTRODUCTION

The ICH Guidelines for Good Clinical Practice discuss and define standards for the development and registration of investigational products. Section 2.5 indicates that *clinical trials should be scientifically sound, and described in a clear, detailed protocol.* The protocol is a document that describes the objective(s), design, methodology, statistical considerations, and organization of a trial[1]. But what does it mean to be scientifically sound as it relates to statistical considerations? While a complete answer would require a lengthy discussion, we focus on statistical considerations as they relate to the number of participants, or sample size, of clinical trials.

The Statistical Principals for Clinical Trials (ICH E9) provides more direct guidance associated with sample size determination in clinical trials. Section 3.5 of E9 states "*The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed.*" This statement is so simple and true, yet the process of determining a sample size is often far more complex. There are typically multiple questions to be addressed, and "large enough" introduces the reality of finite resources (time, money, patients, etc.).

For the remainder of this paper we follow the guidance of E9 that also suggests focus be on the primary objective of a clinical trial. The following additional considerations are also necessary:

- A primary variable
- A test statistic
- Null and Alternative hypotheses
- Type 1 error (probability of erroneously rejecting the null hypothesis)
- Type 2 error (probability of erroneously failing to reject the null hypothesis).

Note that we often refer to statistical power rather than Type 2 error. Power is defined to be 1- Type 2 error, and is typically expressed as a percentage. Methods for sample size determination should be consistent with statistical methods used in the planned analysis, or justification should be provided when this is not the case. Critical assumptions will be made with respect to effect size and variance.

Given the practical perspective of this exercise, most of the remainder of discussion is done via example.

HYPOTHETICAL EXAMPLE

Suppose a pharmaceutical company was developing a product to be used for a medical condition with a high mortality rate. Under standard care, the survival rate after 28 days of treatment could be as low as 20%. The pharmaceutical company believes they may be able to increase this dramatically to almost 80%. In this example, we define:

- The primary variable is a binary response to know if a patient is alive 28 days after starting treatment
- A test statistic appropriate for testing the difference between proportions
- Null hypothesis assumes the proportion of patients alive 28 days after treatment with test product is the same as the proportion alive 28 days after treatment with standard of care. The alternative hypothesis is what we are trying to show, which is that they are different. Actually, we hope the proportion of patients alive 28 days after treatment with test product is larger than the proportion alive 28 days after treatment with standard of care.
- Type 1 error =0.05
- Type 2 error =0.20, suggesting a power of 80%

Interest is really in the percentages of patients that are alive 28 days after initial treatment. This is expressed as:

	Alive after 28 days (Success)	**Not alive after 28 days (Failure)**
Treatment	π_T	1-π_T
Control	π_C	1-π_C

We use this to turn the research question into null and alternative hypotheses. Sometimes it is easier to start by stating the alternative since that should be consistent with what we are trying to prove with the research question, though it is generally written stating the null hypotheses first.

$$H_o:\ \pi_T - \pi_C = 0$$

$$H_a:\ \pi_T - \pi_C \neq 0$$

Note that for this exercise we defined the hypotheses in terms of absolute differences (ie, we subtract one from the other) in success rates. This can be restated in terms of 28-day mortality but we proceed in terms of survival. There are other formulations that are associated with relative risk and odds ratios that are not discussed here.

Also note that the hypotheses are two-sided, meaning the alternative suggests the proportion of patients alive after 28 days on treatment might actually be lower (or even higher) than that on control (standard of care). In actuality we would not be conducting this study if we believed the 28-day survival rate was less than that of standard of care. That of course, would be unethical. In the world of FDA regulations, we often state things as a two-sided hypothesis with α=0.05, which in many cases is equivalent to a one-sided hypothesis with α=0.025

According to the E9 guidance, we need ‘enough’ patients to demonstrate the desired results with a certain degree of confidence. This depends on the actual (ie, true) difference in proportions, a quantity that is unknown. We call this an assumed effect size. Sample size also depends on the variance. When working with proportions, the variance itself is a function of the magnitude of the proportion itself.

Intuitively, one might expect that you need fewer observations to show a difference in proportions if, in fact, that difference was rather large. Conversely, if the proportions were close together but still different, one would need a lot more data to show this.

INITIAL SAMPLE SIZE

A request to calculate a sample size typically comes in the form of stating assumptions about effect size, variance, Type 1 error (alpha), and power (1-Type 2 error). The pharmaceutical company in our example wishes to estimate a sample size with the following assumptions:

π_T	.8
π_C	.2
Alpha	.05
Power	80%

When receiving this request, the statistician in us might want to ask a few more questions. For example, are these numbers realistic? Do you have other data that suggest the assumed percentages? Will the real world even believe the results given our assumptions and sample size?

PROC POWER is the obvious place to start with the initial sample size, though the point-and-click method using SAS Power and Sample Size Calculator is also an option. Given this is an applied exercise, the details of PROC POWER are omitted, thought the reader can easily find them at [3].

Since our interest is in comparing proportions, the TWOSAMPLEFREQ statement is used throughout.

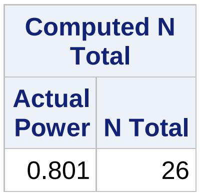
```
proc power ;
   * A two sided approach using GPS option to specify assumed
   proportions.;
   twosamplefreq test=fisher alpha=0.05 gweights=(1 1) gps=(.2 .8)
   power=.8 sides=2 ntotal=.;

   * A two sided approach using GPS option to specify assumed
   proportions.;
   twosamplefreq test=fisher alpha=0.025 gweights=(1 1) gps=(.2 .8)
   power=.8 sides=U ntotal=.;
   * A two sided approach using NULL option to specify assumed
   proportions.;
   twosamplefreq test=fisher alpha=0.05 gweights=(1 1) refproportion=.2
   proportiondiff=.6 power=.8 sides=2 ntotal=.;
run;
```

Sample Code 1

We see multiple TWOSAMPLEFREQ statements in Sample Code 1. This is simply to demonstrate that as with most things in programming, there are multiple ways to obtain a correct answer. The first two statements rely on the GPS option which allows us to simply specify the two assumed proportions. The difference is that the first statement suggests a two sided test at alpha=0.05 whereas the second suggests an upper tailed test at alpha=0.025. The GWEIGHTS suggests a 1:1 allocation between the two groups. The POWER statement allows us to specify the desired power of the study, which is a function of Type 2 error. Recall Power=1-Type 2 error where Type 2 error is failure to reject the null hypothesis given the alternative is true.   Type 2 error results in a missed opportunity, meaning an effective treatment will not be made available to patients in need.

The third approach uses the REFPROPORTION and PROPORTIONDIFF options. Rather than specifying the null and alternative proportions, we specify the reference proportion (.2) and the assumed difference (.6). This approach provides equivalent sample size results as the first two approaches that use the GPS option. There is an advantage to using the third approach in that it easily allows us to specify a range of differences to explore the impact on sample size.

The output of the initial sample size estimation from Proc Power is shown below.



Computed N Total	
Actual Power	N Total
0.801	26

Output 1

Note SAS provides the total n=26 is the minimal sample size needed to achieve the desired power, which suggests n=13 are needed in each group (ie, treatment arm).

Based on the assumptions provided to the statistician, the study team is told that only 13 patients are needed in each group. The first impression is “That’s not enough. No one would believe us. The clinical community says we need at least n=100 for anyone to believe the results.”

N=20

Even though such a sample size may not be viewed as reliable by the clinical community, in the presence of uncertainty about the assumptions the study team asks if there is a way to evaluate the power under different scenarios fixing the total n=20. N=20 is extremely attractive because it greatly reduces the cost of conducting the study. However, in light of the need to satisfy the clinical community, the scenarios for n=20 are simply meant to better understand the impact of assumptions on sample size.

The approach using the REFPROPORTION and PROPORTIONDIFF options easily allow for this. Since it is straightforward to do so (and anticipating future questions), Sample Code 1 was updated to evaluate power under different sample sizes, reference proportions (ie, 28 day survival rate of standard of care), and absolute differences between standard of care and experimental treatment.

```
proc power ;
   twosamplefreq test=fisher alpha=0.05 gweights=(1 1)
   refproportion=.2 .3 .4 .5
   proportiondiff=.2 .25 .3 .35 .4 .45 .5 .55 .6
   power=.8
   sides=2
   ntotal=20 30 40
   power=.;
   plot x=effect;
run;
```

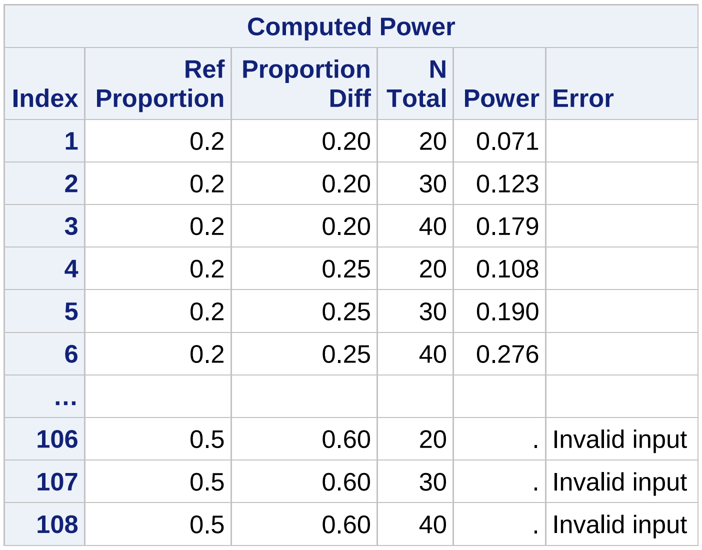
Sample Code 2

Use of REFPRORTION allows us to vary the assumption about the survival rate of standard of care. PROPORTIONDIFF allows us to specify different effect sizes, each to be evaluated at each individual value in the list of reference proportions. For example, REFPROPORTION=.2 in combination of PROPORTIONDIFF=.2 corresponds to a scenario where the survival rate in standard of care is 20% yet the survival rate with experimental treatment is 40%.

Careful readers might note that an assumed reference proportion of .5 with an assumed difference of .6 is invalid. SAS is smart enough to know this, and simply says the combination is invalid in the tabular results.

A list of sample sizes (20 30 40) is specified with the NTOTAL statement, which allows us to be prepared for the eventual question of “what happens if the sample size is 40”.

Since the call to PROC POWER is asking for a large number of power evaluations it produces lengthy tabular output, a fraction of which is shown below in Output 2.



Computed Power

Index	Ref Proportion	Proportion Diff	N Total	Power	Error
1	0.2	0.20	20	0.071	
2	0.2	0.20	30	0.123	
3	0.2	0.20	40	0.179	
4	0.2	0.25	20	0.108	
5	0.2	0.25	30	0.190	
6	0.2	0.25	40	0.276	
…					
106	0.5	0.60	20	.	Invalid input
107	0.5	0.60	30	.	Invalid input
108	0.5	0.60	40	.	Invalid input

Output 2

To aid in presentation of the results, a PLOT statement is used to present a single page figure of the evaluations.

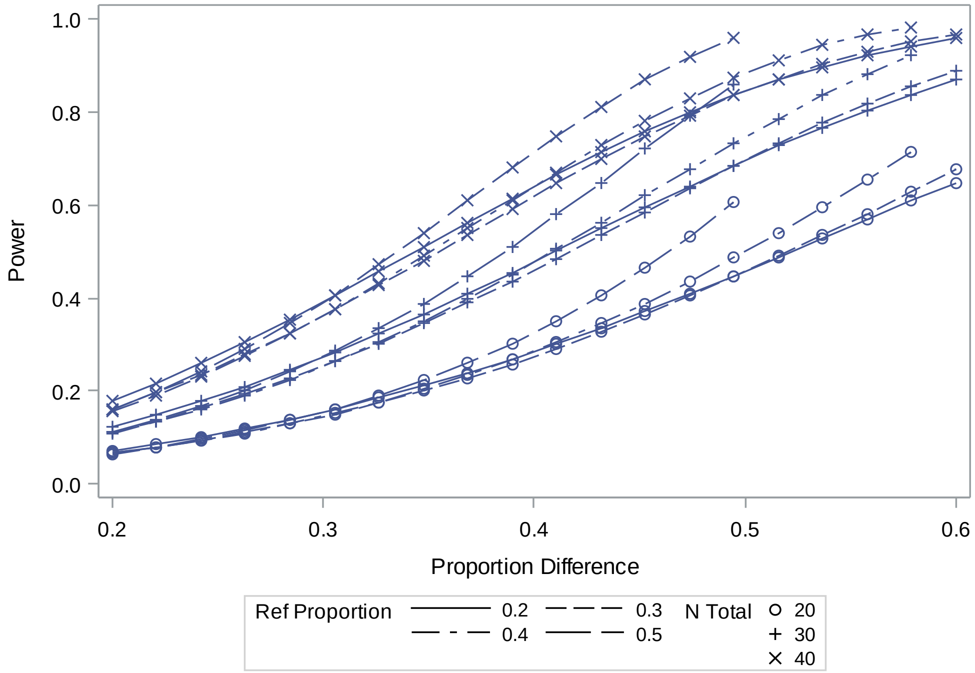


Figure 1

While the output straight from PROC POWER may be readable by a statistician or someone with a trained eye, it would be challenging to report this to a study team. The plot statement can be updated to produce separate plots for each reference proportion by adding VARY(PANEL BY REFPROPORTION). This helps, but the resulting output spans four pages.

Alternatively, an output dataset can be created from PROC POWER using the ODS table OUTPUT. A straightforward update can be made to plot null versus alternative proportions rather than effect size. As seen in Sample Code 3 this was used with PROC SGPLOT to provide the following graph to the study team.

```
data ssn;
   set ssn;
   althyp=refproportion + proportiondiff;
   label althyp='Alternative proportion';
run;

title1 "Power to Assess Decisions Based on Statistical Significance at
N=20, 30, 40";
proc sgpanel data=ssn;
   panelby refproportion;
   series x=althyp y=power/ group=ntotal name='np';
   scatter x=althyp y=power/ group=ntotal;
   keylegend "np" / across=3 position=bottom;
run;
```

Sample Code 3

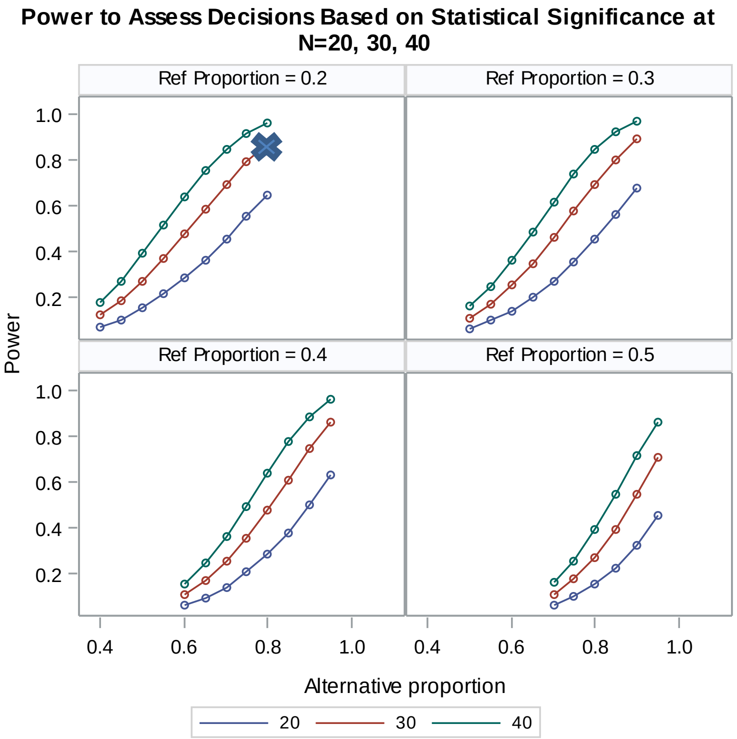


Figure 2

Here we can easily see that the original evaluation of 20% vs 80% can be approximately seen in the first panel where the blue X lies between the n=20 and n=30 lines it the upper left panel. The panel plot easily allows someone to quickly evaluate power under the many scenarios. This helps tremendously in discussions about assumptions during team meetings. In particular, many of the scenarios that are likely to be discussed clearly would have very little power.

RE-EVALUATING ASSUMPTIONS

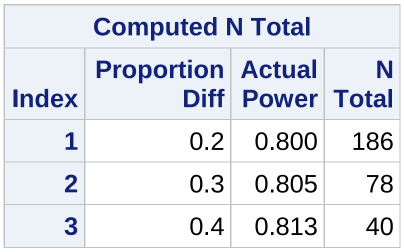
In re-evaluating the assumptions it became clear that the original assumptions of 20% survival in standard of care may have been more suggestive rather than driven by existing data, whether from publication or pilot studies. During team discussions, the study team suggests there is literature stating the survival rate under standard of care is probably closer to 50%.

Sample Code 1 is easily modified to change REFPROPORTION = .2 to REFPROPORTION = .5. In the spirit of anticipating future questions and providing a more complete sample size evaluation, we allowed the difference in proportion to vary from .2 to .4 (by .1). This corresponds to the scenarios where the survival rate for experiment treatment would be 70%, 80%, and 90% while the survival rate of standard of care remains fixed at 50%.

```
proc power ;
   twosamplefreq test=fisher alpha=0.05 gweights=(1 1)
   refproportion=.5
   proportiondiff=.2 .3 .4
   power=.8
   sides=2
   power=.8
   ntotal=.;
run;
```

Sample Code 4

Based on the results shown below (and intuition), the sample size needed to achieve 80% power is substantially higher when there is only a 20% difference in survival rates compared to larger differences.



Computed N Total			
Index	Proportion Diff	Actual Power	N Total
1	0.2	0.800	186
2	0.3	0.805	78
3	0.4	0.813	40

Output 3

While there is some confidence in the assumed survival rate of standard of care, there is no data available and no published literature that provides any information about the survival rate of the experimental treatment. The assumed 80% was purely speculative. The evaluation over different alternative survival rates for experimental treatment proved quite useful.

Given the lack of data, the study team noted that it is very well possible that the survival rate of experimental treatment is 70%, not as optimistic but still clinically important. However, should this be the case, the sample size of n=186 (93 per arm) is too large. Conducting the trial would require more money than the company wished to spend. At this point, a study team member asks “Can’t we start the study, collect some data, and have a look to see how to proceed?” suggesting something called an adaptive design.

CONSIDERATION OF AN ADAPTIVE DESIGN

The use of adaptive study designs has become more common over the years. The ICH Guidelines for Adaptive Designs in Drugs and Biologics provide the following definition: an adaptive design clinical study is defined as a study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study[4].

The guidance provides a lengthy discussion on the use of adaptive designs for studies to be submitted to regulatory authorities. Focus in this exercise is on an adaptive design that allows one or more of the following:

- Stopping a study early due to overwhelming superiority of efficacy
- Stopping a study early due to futility

- Increase in sample size if the original assumptions were to be modified based on early data from within the clinical trial

Group sequential designs (GSD) with appropriate statistical adjustments are considered a generally well-understood adaptive design, meaning the statistical and operational properties are well understood. In the GSD setting, there is an initial commitment to a (slightly larger) sample size with the hope of stopping early, which means they may not need to fully enroll the study leading to reduced costs. The penalty for this is a commitment to a slight increase in sample size that would be required based on the fixed design. Even though there may be a commitment to a slight increase, the hope is that the group sequential methods allow the study to stop early, before the study is fully enrolled. This, in turn, results in potential savings over the fixed design.

To better understand the potential savings, we often refer to an expected sample size. Since some studies will stop early while others will not, the expected sample size is like an average sample size if the study were to be repeated many times. When comparing various study designs, the expected sample size is important to consider when discussing potential savings.

EARLY STOPPING FOR SUPERIOR EFFICACY

PROC SEQDESIGN allows us to evaluate this type of design. Assumptions regarding the 28-day survival remain at 70% for experimental treatment and 50% for standard of care. There is far more to this procedure and its output than can/will be discussed here. For this exercise, we focus on aspects needed for our clinical trial.

```
proc seqdesign pss stopprob errspend ;
      onesidedtwostage: design nstages=2 info=cum(.6 1) alt=upper
      alpha=0.025 beta=.20 method(reject)=errfuncobf;
      samplesize model=twosamplefreq(nullprop=.5 prop=.70 ref=prop
      test=prop);
run;
```

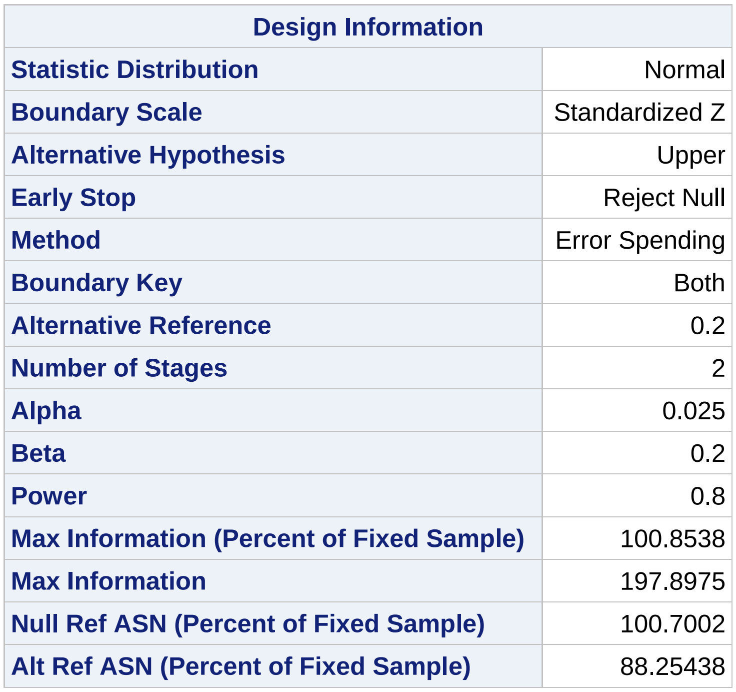
Sample Code 5

As seen the above example, there are two statements needed to evaluate a GSD for our study. The first is the design statement where properties of the design are specified. Recall that the adaptations have to be prospectively defined. ONESIDEDTWOSTAGE is simply a label for the design. NSTAGES=2 indicates our wish to only have 1 interim analysis and 1 final analysis. The first analysis is to be performed once 60% of the statistical information is obtained. Statistical information is closely related to sample size, though they are not equivalent. Fortunately SAS provides the sample size associated with the desired information.

The SAMPLESIZE statement allows us to specify that the test is associated with two sample frequencies, and allows us to specify the null and alternative proportions. Additionally, and mentioned later, the REF statement allows us to specify that the alternative proportions (.7 and .5) be used in generating test statistic rather than the null proportions (.5 and .5). We will see later that using the null proportions (REF=NULLPROP) slightly increases the sample size. This is expected in our example since the variance of a proportion is maximized with the proportion=1/2.

The procedure options PSS, STOPPROB, and ERRSPEND request that the power and sample size summary, stopping probabilities, and cumulative error spending be displayed.

The output from PROC SEQDESIGN starts with a summary of the assumptions:



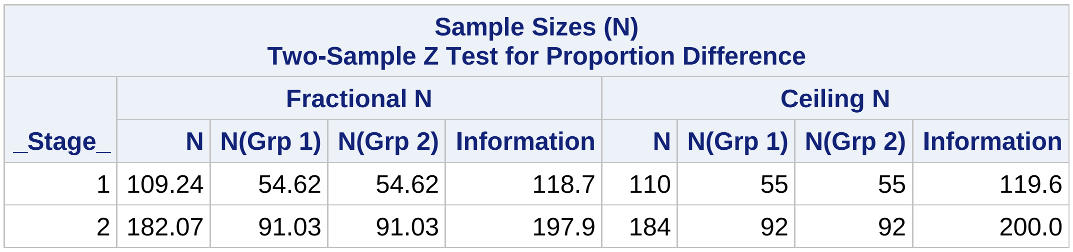
Design Information	
Statistic Distribution	Normal
Boundary Scale	Standardized Z
Alternative Hypothesis	Upper
Early Stop	Reject Null
Method	Error Spending
Boundary Key	Both
Alternative Reference	0.2
Number of Stages	2
Alpha	0.025
Beta	0.2
Power	0.8
Max Information (Percent of Fixed Sample)	100.8538
Max Information	197.8975
Null Ref ASN (Percent of Fixed Sample)	100.7002
Alt Ref ASN (Percent of Fixed Sample)	88.25438

Output 4

The last two rows provide information that compares the sample size from the adaptive design with that from the fixed design. The row for Null Ref ASN suggests the sample size needed for the GSD is 100.7% of that as compared to the fixed design. Thus, a 0.7% increase in sample size is only 2 more patients than would be needed for the fixed design.

Should the alternative hypothesis be true (ie, a 20% difference in the true population proportions), there is a chance for stopping early. This leads to an expected sample size to be smaller than a fixed design. Under the above assumptions, the expected sample size if there is a 20% difference is about 88% that of the fixed design. Fewer patients needed translates to lower costs to conduct the study.

The summary of the sample size is seen below:



Sample Sizes (N)
Two-Sample Z Test for Proportion Difference

	Fractional N				Ceiling N			
Stage	**N**	**N(Grp 1)**	**N(Grp 2)**	**Information**	**N**	**N(Grp 1)**	**N(Grp 2)**	**Information**
1	109.24	54.62	54.62	118.7	110	55	55	119.6
2	182.07	91.03	91.03	197.9	184	92	92	200.0

Output 5

In the call to the procedure, we specified the interim analysis to occur after 60% of the information has been obtained. The output suggests this should occur after the 28-day survival has been obtained from the first 55 patients in each group. Also seen in the summary table is the sample size at the final analysis, which is 92 in each group.

Note that the fixed design discussed earlier suggested n=94 per group while the GSD suggests a fixed design would require slightly less than 100 per group. In the world of sample size determination, different procedures and different software are not expected to yield exactly the same results. Internal algorithms and underlying (asymptotic) assumptions may differ, leading to consistent yet slightly different results.

To assist with evaluation of the design, SAS provides estimates of sample sizes (as a function of the fixed design) under other hypotheses associated with CRef=0.5 and CRef=1.5. This provides a way to approximate sample size if the treatment effect was 50% (and 150%) of that assumed by the alternative. This translates to a difference of 10% or 30% instead of the assumed 20%.

Powers and Expected Sample Sizes
Reference = CRef * (Alt Reference)

		Sample Size
CRef	**Power**	**Percent Fixed-Sample**
0.0000	0.02500	100.7002
0.5000	0.28731	98.5474
1.0000	0.80000	88.2544
1.5000	0.98765	71.5709

Output 6

When dealing with proportions, we know that variance of the test statistic for use in hypothesis testing will depend on the null hypothesis values of the assumed proportions, though the observed sample proportions may be used. In the above example, the proportions from the alternative hypothesis were used in generation of the test statistic since we specified REF=PROP in the TWOSAMPLEFREQ statement. We could have specified the test statistic be based on the null hypothesis via REF=NULLPROP, which results in a slightly larger sample size.

Sample Sizes (N)
Two-Sample Z Test for Proportion Difference

	Fractional N				Ceiling N			
Stage	**N**	**N(Grp 1)**	**N(Grp 2)**	**Information**	**N**	**N(Grp 1)**	**N(Grp 2)**	**Information**
1	118.74	59.37	59.37	118.7	120	60	60	120.0
2	197.90	98.95	98.95	197.9	198	99	99	198.0

Output 7

Given the uncertainty associated with assumptions thus far, the conservative approach is to select the larger sample size for the study design since it is better to have a study that is slightly over-powered than under-powered. A total of 198 patients (99 per group) are needed to achieve at least 80% power.

Boundary Information (Standardized Z Scale)
Null Reference = 0

				Alternative	Boundary Values
	Information Level			Reference	Upper
Stage	**Proportion**	**Actual**	**N**	**Upper**	**Alpha**
1	0.6000	118.7385	118.7385	2.17934	2.66863
2	1.0000	197.8975	197.8975	2.81352	1.98097

Output 8

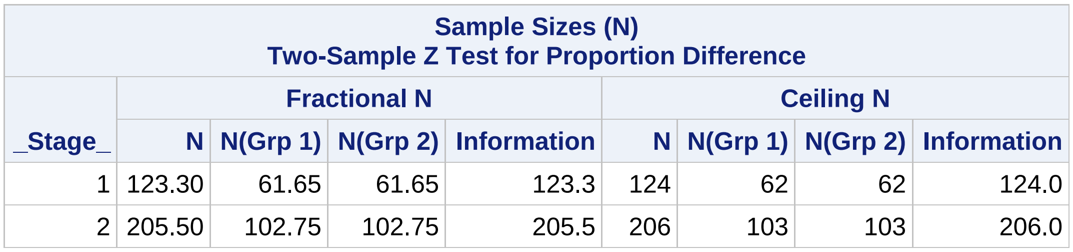
Based on Output 8 the critical values for stopping for efficacy are 2.67 for the interim look and 1.98 for the final look. Should the test statistic at the interim be greater than 2.67, the study may be stopped early for overwhelming efficacy. The boundary scale can easily be changed to provide cutoffs in terms of maximum likelihood estimates or p-values using the BSCALE option on the PROC SEQDESIGN statement.

ADDITION OF EARLY STOPPING FOR FUTILITY

Addition of a futility bound could allow the study to stop early if the likelihood of success after the interim look was extremely low. Intuitively, there should be some evidence of a benefit due to experimental treatment at the interim. If there is none, or the data suggest it is extremely unlikely to have a positive outcome at the conclusion of the study, then it might make sense to stop to study early. This is called stopping for futility.

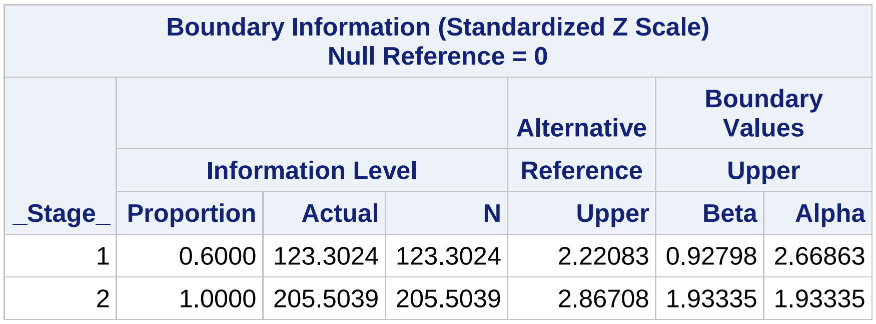
The PROC SEQDESIGN code is easily updated to allow for stopping early for futility at the interim analysis. Sample Code 4 is updated to include two additional options in the design statement: METHOD(ACCEPT)=ERRFUNCOBF STOP=BOTH. This tells the procedure to use the O'Brien Fleming spending function for establishing the futility bound (ie, adjusting to maintain Type 2 error), and that the decision at the interim analysis could be to stop for either efficacy or futility. If the data suggest an ineffective treatment, then the decision may be to stop early and accept the null hypothesis.

The sample sizes and critical values for decision making are shown below:



Sample Sizes (N) Two-Sample Z Test for Proportion Difference								
	Fractional N				Ceiling N			
Stage	N	N(Grp 1)	N(Grp 2)	Information	N	N(Grp 1)	N(Grp 2)	Information
1	123.30	61.65	61.65	123.3	124	62	62	124.0
2	205.50	102.75	102.75	205.5	206	103	103	206.0

Output 9



Boundary Information (Standardized Z Scale) Null Reference = 0						
				Alternative	Boundary Values	
	Information Level			Reference	Upper	
Stage	Proportion	Actual	N	Upper	Beta	Alpha
1	0.6000	123.3024	123.3024	2.22083	0.92798	2.66863
2	1.0000	205.5039	205.5039	2.86708	1.93335	1.93335

Output 10

We see the penalty for adding possible early stopping for futility is a slight increase in sample size. In this case, the sample size required increases to 206. The critical value for stopping early for futility is 0.93. Should the test statistic at the interim analysis be less than 0.93, the study may by stopped early due to futility as it is unlikely to achieve a statistically significant p-value at the conclusion of the study.

It is noted that the final critical values for efficacy differ when the futility bound is added. In fact, the critical value is actually lower than that of the fixed design (1.96) when the futility bound is added. For regulatory reasons, it is proposed that the critical values for early stopping come from a design that does not have a futility boundary so that the critical value for the final analysis is at least as large as the traditional fixed design. The critical values for futility should be added once the critical values for early stopping for efficacy have been obtained. The final sample size should range between the two different study designs. The recommended sample size for this study is 200.

As the study team continues to have reservations about such an up-front commitment, they ask about options to start with a smaller commitment, and potentially increase the sample size after an initial look at the data. This would help by providing better estimates of the survival rates, but it will be based on the existing trial data.

PROMISING ZONE

The typical approach to sample size determination is to first obtain estimates of the treatment effect and variance associated with the primary endpoint from literature or past studies in similar populations. These would be used to obtain an estimated sample size for the desired power. However, there is often uncertainty associated with the assumed estimates of effect size and variance.

This last approach considered is referred to as the Promising Zone, and is thoroughly described in [5]. One of the drawbacks of this approach is that it isn't incorporated in many commonly used software packages. While there is no single SAS procedure, the steps described in [5] can be used with PROC POWER in simulation to obtain expected sample sizes for this adaptive design.

Critical Note: Given this approach is done via simulation in SAS, and this design was not given serious consideration because of operational challenges, the results presented are based solely on that of a single programmer. While the results presented are reasonably consistent with expected results given this design, additional quality control would be implemented if decisions were to actually be made on this design.

As described in [5], the Promising Zone design performs an initial sample size calculation that may be reasonable yet optimistic that yields an initial sample size. An interim analysis is then performed to better understand the underlying assumptions, including the assumed effect size.

At the interim analysis, conditional power is determined based on the observed point estimates. Conditional power can be viewed as the approximate power with the existing sample size if assumptions were based on the point estimates consistent with the observed data. If the observed difference in proportions was larger than the assumed difference, then conditional power would be high, suggesting an over-powered study. Conversely, if the observed difference in proportions was much smaller than the assumed difference, then conditional power would be low, suggesting an under-powered study.

The design classifies the conditional power into 3 zones:

- Zone 1: Unfavorable - low conditional power
- Zone 2: Promising - results are promising, but the power would be lower than expected with the existing sample size
- Zone 3: Favorable - better than expected results

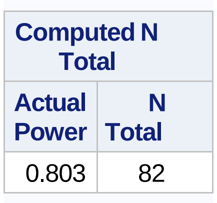
If the conditional power is in Zone 1 or Zone 3, then the decision is to move forward with the study as originally planned with the original sample size. If the conditional power is in Zone 2, then the sample size is re-calculated using the observed point estimates. This results in an increased sample size in hopes of maintaining the desired power.

The cutoffs for the three zones are discussed in general in [5]. Based on [5], if the conditional power was <0.31, then we consider the data as unfavorable and assign it to Zone 1. If the conditional power was above 0.80, then we consider the data as favorable and assign it to Zone 3. All other cases are considered promising, but a sample size adjustment is needed to obtain the desired power of 80%.

According to [5], if the zones are defined using certain criteria then the overall Type 1 error rate is maintained without adjusting the final analysis for the interim look. This allows the usual test statistic to be used rather than an adjusted form to control the Type 1 error rate. Cutoffs for defining Zone 1 under various scenarios are published in [5]. Assuming there is no maximum increase in sample size, we use a conditional power of 31% to define Zone 1.

Implementation

Sample Code 1 was updated with REFPROPORTION=.45 and PROPORTIONDIFF=.3 to obtain an initial sample size assuming the 28-day survival rate for treatment was 75% whereas that of standard of care was 45%. Note that these results are more optimistic about the effect of experimental treatment. The resulting sample size is:



Computed N Total	
Actual Power	N Total
0.803	82

Output 11

This suggests n=41 in each treatment group. While there is an upfront commitment of 82 total patients, the understanding is that a sample size increase may be needed if the results are promising but not as good as originally (and perhaps optimistically) assumed.

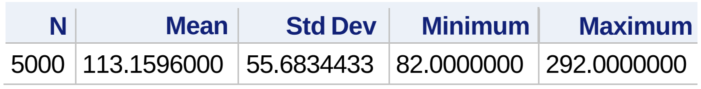
Simulation is required to obtain an expected sample size. Five thousand clinical trials were simulated assuming the true survival rate for standard of care and experimental treatment are 50% and 70%. Note that these assumptions differ from what was used in the initial sample size calculation. Thus, we might expect the conditional power to suggest an increase in sample size. It is also easy to see that a sample size on n=82 is severely underpowered knowing our simulated data differ from the assumptions in the initial power calculation.

Each simulated trial had 200 patients in each treatment group, though not all 200 patients were used.

For each clinical trial, an initial sample of n=21 (~half of the patients) in each group was used to obtain point estimates for the survival rate in each group (and consequently the difference). The point estimates were then used with PROC POWER to obtain a conditional power if the total n remained at 82. The conditional power was placed into 1 of 3 zones. If the individual trial landed in Zone 2, a new sample size was determined based on the observe data. We then increased the sample size for that individual trial in hopes of maintaining the 80% power.

This was repeated for 5000 simulated trials. We were then able to obtain estimates of power in each Zone, and an expected sample size.

With the initial sample of n=21 in each group, the sample size increases in Zone 2 increased the power in Zone 2 from less than 50% to 75%. It did not quite achieve the desired 80% power in Zone 2. As we begin to better understand why, we note that n=21 in each group results in small cell sizes (ie, warnings about expected cell sizes being too small) in many simulated trials. Thinking this could have an impact on distribution assumptions of the test statistic, the initial sample was increased to n=25 in each group. With this small adjustment to the simulations, we were able to achieve the desired 80% power in Zone 2.



N	Mean	Std Dev	Minimum	Maximum
5000	113.1596000	55.6834433	82.0000000	292.0000000

Output 12

Summary statistics were obtained on the total sample size from the 5000 simulated trials.   We see from Output 12 that the expected sample size is n=114. The minimum is the original 82, but some simulated trials suggest the increase can result in a total sample size of 292 which is much larger than the GSD commitment of 200.

We can also look at sample sizes and power broken down by Zone.

Zone	Simulations	Expected Sample Size	Minimum Sample Size	Maximum Sample Size	Power
Zone 1	1840	82	82	82	12%
Zone 2	1924	164	86	292	80%
Zone 3	1236	82	82	82	91%

Output 13

We see the power associate with Zone 2 is maintained at 80%, much larger than the approximate 50% power in Zone 2 should we not have increased the sample size.

The design parameters for defining Zone 1 could have easily been updated to incorporate a maximum sample size increase. However, that was not considered for this exercise. Also not considered was verification of maintaining Type 1 error. The simulation could be easily modified to show that the design parameter used do not inflate the overall Type 1 error.

Upon presenting the results to the study team, there were two major areas of concern. First was associated with the guidance on the use of adaptive designs. The Promising Zone design is not considered as a design that is generally well-understood. The mathematics are understood, but the operational characteristics are not. One major criticism is that the observed treatment difference at the interim could be inferred from the magnitude of the increase in sample size. Steps would need to be taken to prevent this. The second major area of concern was with the potential for the sample size to increase to almost 300. If this design were to receive more serious considerations, we would want to modify the simulations to account for some pre-determined maximum allowable increase.

CONCLUSION

Determining a sample size is a process that can evolve as demonstrated in the above hypothetical example. Statistical considerations, operational challenges, and financial constraints all play an important role. It often takes in-person meetings with the statistician and study team members to fully vet the assumptions that go into selecting a sample size, and it is often an iterative exercise.

The study team needs to fully understand operational considerations of an adaptive design. The most important aspect to consider is that the interim analysis is to be performed by a group independent of the study team, and ideally independent of the sponsor organization. The pharmaceutical company should not get access to any interim results unless the decision is to stop early. This is a concept that many non-statisticians don't fully understand or appreciate, and can lead to a hesitation to even consider the adaptive design approach. Nevertheless, it is a fundamental constraint that must be considered to maintain integrity of the study results.

SAS provides a number of procedures to assist with sample size determination in most commonly used designs. As the industry moves toward other less well understood adaptive designs, simulation in SAS will be required to evaluate power and sample size.

REFERENCES

[1] ICH GCP Guidelines

[2[E9 Statistical Principles for Clinical Trials

[3] http://support.sas.com/documentation/cdl/en/statug/68162/HTML/
default/viewer.htm#statug_power_overview.htm

[4] https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf

[5] Mheta, C, Pocock, S., 2000, “Adaptive Increase in Sample Size when Interim Results are Promising: A Practical Guide with Examples”, Statistics in Medicine, 00:1-6

ACKNOWLEDGMENTS

The author would like to thank Amber Randall, David Kerr, Marcia Craib and Darrin Despain for their careful review of this paper. The author would also like to acknowledge the many programmers publishing information on the internet, and apologizes to those not properly referenced above, citing the inherent nature of internet searching which makes it difficult to track so many individual ideas.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

William Coar, PhD
Biostatistician/Director, Statistical Consulting
Axio Research, LLC
Seattle, WA 98121
Email: williamc@axioresearch.com