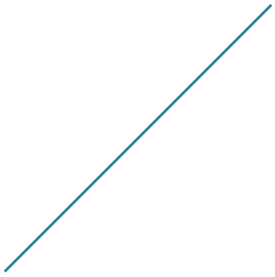
Generic Placebo Controlled Study -Two group Sample Size Determination

Prepared for:

Prepared by:

Proposal number: 000-001



GENERIC

EXECUTIVE SUMMARY

Objective

Determine a sample size which ensure the Axial experiment will have statistical significance, enough power and reasonable effect size. A _____% buffer should probably be included to account for any unaccounted sources of variation.

Outline

All calculations will be demonstrated below and are subject to review, to ensure consistency and accuracy.
The Mini Axial Experiment being performed at _____ serves as the basis for the data and calculations performed below. The primary researchers for that experiment are _____ and _____. _____ oversaw the experiment and is acting Principal Investigator.

The experiments are still ongoing as of _____, however all the data used and summarized below comes form experiments on only 2 days, namely _____ and _____. Prior experiments lacked the accuracy of a jig that could _____ the samples, in a controlled, consistent and repeatable fashion. Subsequent experiments performed in March of the same year suffered from failed positive controls, indicating that our stored bacteria were dying. Future data can and should be incorporated into the sample size calculations outlined below. It should be noted that even incorporating the data from the prior experiments that were less consistent, the mean and SD only varied by less than a factor of 2.

The rest of this paper will outline the derivation and method of the sample size calculation, leaving us with 4 key variables. These are

Number of Groups - self explanatory
Effect Size - A measure of the magnitude of the phenomenon
Confidence Level - A desired frequency of confidence intervals that contain the true estimate of the parameter in question for the observed data (usually 95%)
Power - The probability of rejecting the null hypothesis when the alternative hypothesis is true. It is a function of the probability of making a Type II Error. (False negative).

GENERIC

AXIAL RESULTS

Raw Data

Sample Number (0 seconds)	CFU Count
1	3.25E+04
2	5.50E+04
3	4.00E+04
4	6.25E+04
5	6.00E+04
6	3.50E+04
7	3.00E+04
8	4.00E+04
9	1.25E+04
10	2.25E+04
11	1.00E+04
12	1.50E+04
13	1.00E+04
14	5.00E+03
15	2.00E+04
16	3.00E+04
17	3.25E+04
18	1.00E+04

Mean:29027.77778

SD:17680.5578

Sample Number(87 seconds)	CFU Count
1	0.00E+00
2	0.00E+00
3	0.00E+00
4	0.00E+00
5	0.00E+00
6	2.50E+01
7	2.50E+01
8	0.00E+00
9	1.25E+02
10	2.00E+02
11	2.25E+02
12	2.50E+02
13	3.00E+02
14	4.00E+02

Mean:96.875

SD:134.1252524

GENERIC

DERIVATIONS

For the purposes of this study, no initial assumptions are made about the shape of the distribution, of the interest groups or control groups. .

Let cv be the critical value and α be the probability of rejecting the null hypothesis H_0.

$$\begin{aligned}
\alpha &= Pr(X \geq cv \,|\, H_0) && \text{Definition of Type I error} \\
&= 1 - Pr(X \leq cv \,|\, H_0) && \text{Probabilities sum to 1} \\
&= 1 - Pr\left(\frac{X-\mu_0}{\sigma_n} \leq \frac{cv-\mu_0}{\sigma_n} \,\middle|\, H_0\right) && \text{Standardize both sides of the inequality} \\
&= 1 - \Phi\left(\frac{cv-\mu_0}{\sigma_n}\right) && \text{Definition of standard normal distribution function} \\
1-\alpha &= \Phi\left(\frac{cv-\mu_0}{\sigma_n}\right) && \text{A little algebra} \\
z_{1-\alpha} &= \frac{cv-\mu_0}{\sigma_n} && \text{Definition of standard normal quantiles} \\
cv &= \mu_0 + z_{1-\alpha}\,\sigma_n && \text{A little more algebra}
\end{aligned}$$

With clear values established for when to accept the null hypothesis, we can now derive the power formula.

$$\begin{aligned}
\text{Power} &= Pr(X \geq \mu_0 + z_{1-\alpha}\,\sigma_n \,|\, H_1) && \text{Definition of power} \\
&= 1 - Pr(X \leq \mu_0 + z_{1-\alpha}\,\sigma_n \,|\, H_1) && \text{Probabilities sum to 1} \\
&= 1 - Pr\left(\frac{X-\mu}{\sigma_n} \leq \frac{\mu_0 + z_{1-\alpha}\,\sigma_n - \mu}{\sigma_n} \,\middle|\, H_1\right) && \text{Standardize both sides of the inequality} \\
&= 1 - Pr\left(\frac{X-\mu}{\sigma_n} \leq \frac{\mu_0-\mu}{\sigma_n} + z_{1-\alpha} \,\middle|\, H_1\right) && \text{A little algebra} \\
&= 1 - \Phi\left(\frac{\mu_0-\mu}{\sigma_n} + z_{1-\alpha}\right) && \text{Definition of standard normal distribution function} \\
&= \Phi\left(\frac{\mu-\mu_0}{\sigma_n} - z_{1-\alpha}\right) && 1-\Phi(x) = \Phi(-x)\;;\; z_a = -z_{1-a}
\end{aligned}$$

We can now use al alternative definition of power to calculate sample size, namely Power = $1-\beta$.

$$1-\beta = \Phi\left(\frac{\mu-\mu_0}{\sigma_n} - z_{1-\alpha}\right) \qquad \text{Power formula from above}$$

$$z_{1-\beta} = \frac{\mu-\mu_0}{\sigma_n} - z_{1-\alpha} \qquad \text{Definition of standard normal quantiles}$$

$$\frac{1}{\sigma_n} = \frac{z_{1-\beta}+z_{1-\alpha}}{\mu-\mu_0} \qquad \text{A little algebra}$$

Now that we have derived the sample size, we assume that both the not inoculated group and the control group should have similar means and distributions of bacteria. In addition we are assuming that the data is well simulated by a normal distribution(randomness via Central Limit Theorem). Let n_g be the number of groups and let every subgroup be i.i.d such that $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

The expected value of the mean is just the mathematical sum of all the means, divided by n and we have that

$$\bar{Y} \sim N(\mu, \sigma^2/n)\,.$$

And if we have multiple subgroups,

$$\bar{Y} \sim N(\mu, \sigma^2/(n/n_g))\,.$$

Substituting σ_n with $\sigma/\sqrt{n/n_g}$ we find the required power and sample size required below.

$$\text{Power} = n_g \cdot \Phi\left(\frac{\mu-\mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha}\right)$$

$$n = n_g\left(\sigma\frac{z_{1-\beta}+z_{1-\alpha}}{\mu-\mu_0}\right)^2$$

Let Effect Size $E_s = \dfrac{(\mu-\mu_0)}{\sigma}$

$$n = n_g\left(\frac{z_{1-\beta}+z_{1-\alpha}}{E_s}\right)^2$$

The idea is to split the study up into multiple subgroups, where one subgroup is given a treatment and the other subgroup is not given a treatment(control group). The goal is to detect a meaningful reduction in bacteria count. Due to the mini axial experiment performed at _____(results above), we are able to reasonably estimate the mean and variance of both the target and the known group. These preliminary results are relied upon to estimate a sample size for the Axial Experiment.

***Note that this study design is analogous to a placebo-controlled study where a selected group of patients is divided randomly into two groups. One group is given a placebo(control) and one group is given the treatment.

Hypothesis, Null Hypothesis and Two Sided Test

Based on the study design, the null hypothesis would be that the _____ does not result in a significant reduction(log 4 for instance) in bacteria counts.
A two sided test, which rejects both types of extreme results does not appear to be obviously superior. Overdosage with _____ which would result in little to no bacteria detected, is not harmful to the patient and is only negative in the context of the _____ and the stress that the power drain places upon the _____.

Thus, a single sided test which rejects the the null hypothesis, only too much bacteria is detected on the treated group seems to make more sense. Nonetheless, calculations for both of these test are presented below. As you will see below, the discrete nature of samples sizes results in the same ultimate group size.

Power, Effect Size and Magnitude

One of the most commonly used measures of statistical significance is a p-value of <0.05. Significance however, says little about the magnitude of the effect between a control group and a treatment group. In contrast, effect size is quantitative measure of the magnitude of a given phenomenon. Effect sizes are commonly reported and scrutinized in published studies and we will seek to use a few common effect sizes which measure the differences in mean values between the two groups. Three common used measures of effect size are discussed here - Cohen's D, Glass's Delta and Hedge's G.

Cohen's D is simply the mean difference, over the variance $\frac{|(\mu - \mu_0)|}{\sigma}$ which as we know from above, is a quantity that we encountered in our power size calculation. This statistic makes most sense when the variance of both groups is the same. Unfortunately, based on the _____ Mini Axial Experiment data, we know this not to be the case. The variance of the untreated group is orders of magnitude larger than the treated group.
Thus, we are left with the option of using the variance of the control group which is known.
Glass's Delta has the same definition, except that the standard deviation used is the standard deviation of the second group. (treatment group)
Pooling the variance to find an aggregate is another possible alternative. Hedge's G defines the pooled variance as the weighted average of the variance, based on the sample size of each group.

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)^2 s_k^2}{n_1 + n_2 + \ldots + n_k - k}$$

Calculations for all these various measure of effect size will be outlined below.

For a primer on effect size vs p-value, suggested reading: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/

Sample Size Determination

$n = n_g \left(\frac{z_{1-\beta} + z_{1-\alpha}}{E_s} \right)^2$ represents the needed sample size, based on:

n_g — the number of groups in the study

$Z_{1-\beta}$ — The power of the study (Note: Beta is the probability of making a Type II error)

$Z_{1-\alpha}$ — The size of the desired Confidence Interval

E_s — The Effect Size

From inspection, one can see that as the number of groups increases, the number of required samples increases linearly as well, thus, we can calculate the sample size for just 2 groups and scale that number accordingly to work well with n groups for instance, assuming that the groups in question have similar variances and means. Otherwise, it the effect size needs to be recalculated.

As discussed above, for a two sided test, we actually use the same Z value for power but the Confidence Interval Value of $Z_{1-\alpha/2}$. (Note that these are values from a standard normal table). There

We will assume a 95% Confidence Interval and a power of 90%.

As listed above —
Control Group Mean:29027.77778
Control Group SD:17680.5578
87 Second Group Mean: 96.875
87 Second Group SD: 134.1252524

Using these numbers, we find that

Cohen's d = (96.875 - 29027.77)/12502.396528 = 2.314028

Gates' delta = (96.875 - 29027.77)/17680.55 = 1.636312

Hedges' g = (96.875 - 29027.77)/13309.727377 = 2.173665

	90%	95%	99%	99.9%
$Z_{1-\alpha}$	1.28	1.65	2.33	3.29
$Z_{1-\alpha/2}$	1.65	1.96	2.58	3.09

Using the values from the standard normal table shown above —

A confidence interval of 95% corresponds to a $Z_{1-\alpha}$ value of :1.65
A confidence interval of 95% corresponds to a $Z_{1-\alpha/2}$ value of :1.96
A power level of 99% corresponds to a $Z_{1-\beta}$ value of : 2.33

Using the most conservative effect size, namely Gates' Delta, a 95% two sided confidence interval and a 90% power, substituting into $n = n_g\left(\frac{z_{1-\beta} + z_{1-\alpha/2}}{E_s}\right)^2 = 2(\frac{1.96 + 1.28}{1.636312})^2 = 7.8413 \approx 8$

For a one sided test with Gates' Delta, substituting into
$n = n_g\left(\frac{z_{1-\beta} + z_{1-\alpha}}{E_s}\right)^2 = 2(\frac{1.65 + 1.28}{1.636312})^2 = 6.4125 \approx 8$, so that we have similar sized groups.

In fact, if we were to use the most conservative 99.9% values,
$n = n_g\left(\frac{z_{1-\beta} + z_{1-\alpha/2}}{E_s}\right)^2 = 2(\frac{3.29 + 3.09}{1.636312})^2 = 30.4046 \approx 32$ so that we have similar sized groups.

Thus, for a power of **90% and a CI of 95%, we need 4 samples per ground**. If we wish to raise that to **99.9% power and 99.9% CI, we need 16 samples per group**.

Challenges and Assumptions in applying this to the Non Clinical

The biggest challenge is not knowing the variance of the other groups, namely of _____. It seems reasonable to assume that _____ sterilization is less likely to result in such a low mean bacteria count. The raw data from the _____ mini-axial demonstrated the plate counts were often in the single digit range, below the which lowers the mean variance (standard error) because of the lack of deviation from zero.
Based on the effect sizes, that we chose, this only becomes a problem if the variance and therefore SD of the unknown groups are larger than the control group. This does not appear to be a likely outcome as it is not unreasonable to expect at least a log reduction of bacteria from the ethanol caps, since _____ are an FDA-clinically approved product. However, considering the time and expense needed to conduct the Axial study, it may be wise to choose a larger sample size, significantly larger than 4 per group if resources are permitting, to allow us a buffer in the event of unexpectedly large variance in one of the test groups.

Recommended sample size: 16 per group (99.9/99.9) or 4 per group(90/95)

Once rough cost and time estimates are available, we can use Prior Probabilities — Bayes Law to evaluate the feasibility of conducting subsequent trials in the event of negative results using 4 per group.