Determining sample size for a completely randomized experiment to achieve an acceptable power with a Shiny app

true

Last edit: November 25, 2024

Abstract

Summary Understanding and computing power and the relationship between sample size and power are facilitated via a Shiny app. The Shiny app allows students to solve various scenarios by entering different parameters or moving sliders.

Keywords: Power, Shiny app, significance level, type I error, type II error

INTRODUCTION

The concept of power and its relationship to sample size for a single sample via a Shiny app was discussed in Arnholt (2019). This work shows how the noncentrality parameter (λ) is a measure of statistical difference between **population** means and the F value computed in an ANOVA table is a measure of statistical difference between **sample** means. A Shiny app is presented where students can experiment with different design structures (i.e. different sample sizes for each of the a treatments) to ensure their experiments attain satisfactory power. The notation used for the one-way completely randomized design follows that presented in Ugarte, Militino, & Arnholt (2015). To aid with the connection between noncentrality parameters and test statistics, it is shown how the pooled variance t-test is a special case of the F-test when there are ($a = 2$) treatments before generalizing to $a >= 2$ treatments.

STATISTICAL BACKGROUND

The observations in a completely randomized design (CRD) can be described with a linear statistical model

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \text{ for } i = 1, 2, \ldots, a \text{ and } j = 1, 2, \ldots, n_a \tag{1}$$

where Y_{ij} is the j^{th} observation of the i^{th} treatment, μ is a parameter common to all treatments called the overall mean, τ_i is a parameter unique to the i^{th} treatment called the i^{th} treatment effect, and ϵ_{ij} is a random error component. For hypothesis testing, the model errors are assumed to be normally and independently distributed with mean zero and constant standard deviation $\big(NID(0, \sigma)\big)$. Although estimating the parameters for Model (1) is possible, the goal of the experimenter is typically to discern whether or not the a treatment means are equal. The hypotheses of interest are

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \quad \text{versus} \quad H_1 : \mu_i \neq \mu_j \text{ for some } (i, j).$$

The notation that follows is adopted from Ugarte et al. (2015). The sum of the observations in the i^{th} treatment group is $Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}$, and the mean of the observations in the i^{th} treatment group is $\bar{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = \frac{Y_{i\bullet}}{n_i}$. The bar indicates a mean while the dot ($\bullet$) indicates that values have been added over the indicated subscript. The sum of all observations is $Y_{\bullet\bullet} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}$. The grand mean of all observations is denoted $\bar{Y}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij} = \frac{Y_{\bullet\bullet}}{N}$. For $a = 2$ treatments, the typical sum of squares for testing the hypotheses in () are shown to the left of the equivalence ($\equiv$) symbol in (2). The

representation on the right side of the $\equiv$ in (2) is the standard test statistic used to test two means when variances are assumed to be equal. To verify the equivalence keep in mind that $\sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_{i\bullet} n_i$ for $i = 1, 2$ and $(n_1 + n_2)\bar{Y}_{\bullet\bullet} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}$.

$$F = \frac{MS_{\text{Treatment}}}{MS_{\text{Error}}} = \frac{\dfrac{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}{df_{\text{Treatment}}}}{\dfrac{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2}{df_{\text{Error}}}} \equiv \frac{(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{S_p^2} \tag{2}$$

Rewriting the right side of (2) yields

$$F = \frac{(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})^2}{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \left[\frac{(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right]^2 = [t]^2. \tag{3}$$

The quantity in (2) measures the statistical differences between **sample** means. Replacing the sample means in (2) with the population means yields

$$\lambda = \frac{(\mu_1 - \mu_2)^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{\sigma^2} = \frac{SS_{\text{Hypothesis}}(\text{population})}{\sigma^2}$$

where $SS_{\text{Hypothesis}}(\text{population})$ is the sum of squares for treatments obtained by replacing $\bar{Y}_{1\bullet}$ with μ_1, $\bar{Y}_{2\bullet}$ with μ_2, and $\bar{Y}_{\bullet\bullet}$ with $\frac{n_1\mu_1 + n_2\mu_2}{n_1 + n_2}$.

EXAMPLE

The following scenario can be given to students:

Consider a ficticious experiment where one of two hormones (testosterone or isoandrostenolone) are given to day old male chicks for fifteen days. At the end of the fifteen days, the experimenter hypothesizes the average weight of chicks that receive testosterone will be 100 mg and the average weight of chicks that receive isoandrostenolone will be 70 mg. Based on previous work, the researcher estimates the standard deviation for both groups of chicks to be somewhere between 20 mg and 30 mg. What is the minimum number of chicks that should be assigned to each group to obtain a power for the test of at least 0.80 using $\alpha = 0.05$?

Solution using base R functions

Start by asking the students to specify the null and alternative hypotheses they will use to test the two hormones. The hypotheses are:

$$H_0 : \mu_{\text{testosterone}} = \mu_{\text{isoandrostenolone}} \quad \text{versus} \quad \mu_{\text{testosterone}} \neq \mu_{\text{isoandrostenolone}}.$$

Since the researcher specified a range for σ, two sets of values will be computed one for when the value of σ is 20 mg and one for when the value of σ is 30 mg. Start by finding a critical value for which the null hypothesis will be rejected at the $\alpha = 0.05$ level with a guess of $n_1 = n_2 = 10$. That is we are looking for $F_{0.95,2-1,10+10-2} = F_{0.95,1,18} = 4.4138734$. Next compute the noncentrality parameter when $\sigma = 20$ mg.

$$\lambda = \frac{(\mu_1 - \mu_2)^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{\sigma^2} = \frac{SS_{\text{Hypothesis}}(\text{population})}{\sigma^2} = \frac{\sum_{1=1}^{a} n_i(\mu_{i\bullet} - \bar{\mu}_{\bullet\bullet})^2}{\sigma^2} = \frac{10(100 - 85)^2 + 10(70 - 85)^2}{20^2} = 11.25$$

$$\text{Power}(\lambda = 11.25) = P(F^*_{1,18,\lambda=11.25} \geq f_{0.95,1,18} = 4.4138734) = 0.8869702$$

Since we only need to achieve a power of 0.80, we can reduce the values for each sample. Next, we consider using $n_1 = 9$ and $n_2 = 8$ when $\sigma = 20$ mg.

$$\lambda = \frac{(\mu_1 - \mu_2)^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{\sigma^2} = \frac{SS_{\text{Hypothesis}}(\text{population})}{\sigma^2} = \frac{\sum_{1=1}^{a} n_i(\mu_{i\bullet} - \bar{\mu}_{\bullet\bullet})^2}{\sigma^2} = \frac{9(100-85)^2 + 8(70-85)^2}{20^2} = 9.5625$$

$$\text{Power}(\lambda = 9.5625) = P(F^*_{1,15,\lambda=9.5625} \geq f_{0.95,1,15} = 4.5430772) = 0.8236915$$

```r
# Hypothesized means
hypmeans <- c(100, 70)
# Number of treatments
a <- length(hypmeans)
n1 = 10
n2 = 10
# Total number of experimental units
N <- n1 + n2
# Degrees of freedom for error
dferror <- N - a
# Sigma value
sigma <- 20
# Create n1 values of 100 and n2 values of 70 and store in Y
Y <- rep(hypmeans, times = c(n1, n2))
# Create a treatment factor with n1 values of testosterone and n2 values of isoandrostenolone
Treat <- factor(rep(c("testosterone", "isoandrostenolone"), times = c(n1, n2)))
# Compute SS for ANOVA
summary(aov(Y ~ Treat))
```

```
            Df Sum Sq Mean Sq   F value Pr(>F)
Treat        1   4500    4500 1.993e+31 <2e-16 ***
Residuals   18      0       0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Pull out the SS Treat value and assign to SShyp
(summary(aov(Y ~ Treat))[[1]][1, 2] -> SShyp)
```

```
[1] 4500
```

```r
# Noncentrality parameter
(lambda <- SShyp/sigma^2)
```

```
[1] 11.25
```

```r
(CriticalF <- qf(0.95, a - 1, N - a))
```

```
[1] 4.413873
```

```r
# Power for lambda
(pf(CriticalF, a-1, dferror, lambda, lower = FALSE) -> POWER)
```

```
[1] 0.8869702
```

Solution using the Shiny app with a *t*-test

A more conservative estimate of the required sample size is obtained by using a t-test. The symmetric curve on the right side of Figure depicts a central t-distribution with $n-1$ degrees of freedom. The slightly left skewed curve on the left side of Figure is a non-central t-distribution. To find the required sample size with a t-test, the amount of asymmetry in the non-central t-distribution is computed by the Shiny app when the user provides a best guess for the population standard deviation (σ) in conjunction with the input values for μ_0 and μ_1. In this scenario, suppose 50 mg/dL is an estimate of the standard deviation for the population of interest in the problem. Therefore, 50 mg/dL is used as the best guess for the population standard deviation. When using a t-test to compute the required sample size, click the Variance unknown (t-test) radio button. The rest of the buttons and values with the exception of the Sample size (n) : box will be the same as those used for the z-test. Increase the value of n until the power is at least 0.80. Using $n = 43$ returns a power value of 0.8012 as shown in Figure .

##Step by step solution with a z-test {-}

A step by step solution should start by asking the students to specify the null and alternative hypotheses they will use to test that the new drug reduces cholesterol. Although the null and alternative hypotheses for this scenario are generally written as:

$$H_0 : \mu = 300 \text{ versus } H_A : \mu < 300,$$

it helps to remind the student that in using a z-test, one is also assuming the distribution of the quantity in question (cholesterol) has a normal distribution with a known mean and known standard deviation of 300 mg/dL and 50 mg/dL, respectively. Consequently, the sampling distribution of the sample mean ($\bar{X}$) has a known normal distribution with a mean of 300 mg/dL and a standard deviation (or standard error of $\bar{X}$) of $50/\sqrt{n}$. Have the students sketch by hand a graph like the one in Figure, shading the quantities α, β, and Power when the true mean is 275. After students sketch Figure, have them translate the English prose of the desired average reduction in cholesterol to a statement such as : The power of the test if the true distribution is centered at 275 with a standard deviation of $50/\sqrt{n}$ is the probability the null hypothesis will be rejected when $\alpha = 0.01$. Ask the students to solve for the number of subjects (n) they should include in their sample when using a z-test so that Power $\left(\mu_1 = 275\right) \geq 0.80$. The null hypothesis when $\alpha = 0.01$ for the alternative hypothesis $H_A : \mu < \mu_0$ will be rejected for standardized test statistic values less than or equal to $z_{0.01} = -2.3263$, where the num in z_{num} denotes the area to the left of a standard normal distribution. That is, the null hypothesis $H_0 : \mu = 300$ will be rejected when the sample mean, $\bar{X}$, is less than $300 - 2.3263 \cdot 50/\sqrt{n}$. When $\mu_1 = 275$, the power is

$$\text{Power}\left(\mu_1 = 275\right) = P\Big(\bar{X} \leq 300 - 2.3263 \cdot \tfrac{50}{\sqrt{n}} \text{ given } \bar{X} \text{ has a mean of } 275\Big)$$

To each side of the inequality

$$\bar{X} \leq 300 - 2.3263 \cdot \tfrac{50}{\sqrt{n}}$$

subtract the mean (275) and divide by the standard deviation ($50/\sqrt{n}$) to obtain a random variable that follows a standard normal distribution (z).

$$\begin{aligned}
\text{Power}\left(\mu_1 = 275\right) &= P\left(\frac{\bar{X} - 275}{\frac{50}{\sqrt{n}}} \leq \frac{(300 - 275) - 2.3263\frac{50}{\sqrt{n}}}{\frac{50}{\sqrt{n}}}\right) \\
&= P\left(z \leq \frac{\sqrt{n}}{2} - 2.3263\right)
\end{aligned}$$

Since the problem requires Power $\left(\mu_1 = 275\right) \geq 0.80$, the quantile from a standard normal distribution with 0.80 of the area to the left, $z_{0.80} = 0.8416$ is used to solve for n below. The value of $z_{0.80}$ can be found with software or using a standard normal table.

$$0.8416 \leq \frac{\sqrt{n}}{2} - 2.3263$$
$$n \geq 40.1424$$

To achieve a power of at least 0.80, one needs to use 41 subjects. Note that one will always round to the next largest integer to achieve the required power specified in the problem.

After the students solve the problem, have them verify their answer with the Shiny app at https://mathr.math.appstate.edu/shiny/Power2/. Students will specify the alternative hypothesis ($H_A : \mu < \mu_0$), select the appropriate test (z-test), enter the hypothesized mean (300), enter the true mean (275), type the population standard deviation (50), select the significance level (0.01), and then experiment with increasing or decreasing the sample size to find the minimum value of n that achieves a power greater than or equal to 0.80 as shown in the title of the graph in Figure **??**.

Step by step solution with a *t*-test

The step by step solution with a t-test is similar to the solution with a z-test. Students specify their hypotheses and sketch by hand a graph like the one in Figure shading the quantities α, β, and Power $(\mu_1 = 275)$. Only after sketching Figure , have the students translate the English prose of the desired average reduction in total cholesterol to a mathematical statement such as .

$$\text{Power}\left(\mu_1 = 275\right) = P\big(\text{Reject } H_0 \text{ given } H_A \text{ is true}\big) \tag{4}$$

A more conservative estimate of the required sample size is obtained when one uses the t-test. When the null hypothesis is true, the quantity in (5)

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \tag{5}$$

follows a central t-distribution, generally denoted as t_{n-1}, where $n-1$ are the degrees of freedom for the distribution. When the null hypothesis is false, the quantity in (5) follows a non-central t-distribution, generally denoted as $t_{n-1;\gamma}$, where $n-1$ are the degrees of freedom for the distribution, and γ is referred to as the non-centrality parameter, a measure of the asymmetry of the distribution. To find the required sample size with a t-test, first determine the non-centrality parameter by using a best guess for σ. Unfortunately, the process is now more complex as the non-centrality parameter gamma (γ) is a function of n,

$$\gamma = \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{275 - 300}{\frac{50}{\sqrt{n}}} = \frac{-\sqrt{n}}{2}.$$

Finding the power when $\mu_1 = 275$ with a t-test is computed by finding the probability of rejecting the null hypothesis given $\mu_1 = 275$ (H_A is true). Recall that the null hypothesis is rejected for values to the left of a central t distribution with 0.01 in its left tail and $n-1$ degrees of freedom, denoted as $t_{0.01;n-1}$. For example, $t_{0.01;40} = -2.4233$, can be found using either a table or with statistical software. Consequently, the power when $\mu_1 = 275$ is the area in a the non-central t-distribution ($t_{n-1;\gamma}$) to the left of $t_{0.01;n-1}$, written mathematically as:

$$\text{Power}\left(\mu_1 = 275\right) = P\Big(t_{n-1;-\sqrt{n}/2} \leq t_{0.01;n-1}\Big),$$

and depicted as the blue and purple shaded areas in Figure **??**.

To find a solution, one will need to use increasing values of n until the $P\big(t_{n-1;-\sqrt{n}/2} \leq t_{0.01;n-1}\big) \geq 0.80$. Since the solution for the z-test yielded an $n = 41$, it seems reasonable to start with that value. Although

tables and charts for the non-central t-distribution are available, most tables have limited values of γ, and most charts require extensive interpolation. Consequently, one should use a software program that will work with non-central distributions. To find $P\left(t_{n-1;-\sqrt{n}/2} \leq t_{0.01;n-1}\right)$, one solution using `R` is given next. The `qt()` and `pt()` functions in `R` return a t quantile and the area to the left of a quantile in a t-distribution, respectively. The first two arguments for `qt()` are the area to the left of the desired quantile and the degrees of freedom. The first three arguments for `pt()` are the quantile, the degrees of freedom, and the non-centrality parameter. The critical value of t with $n = 41$ is `qt(0.01, 41 - 1)`, stored in `t1`. Note that text following the pound sign (`#`) in `R` are comments.

```
t1 <- qt(0.01, 41 - 1)  # Critical t value, alpha = 0.01, dof = 40
t1
```

```
[1] -2.423257
```

The area to the left of `t1` in a non-central t distribution with $n = 41$ and $\gamma = -\sqrt{41}/2$ is the power and is computed with the command `pt()` shown below.

```
pt(t1, 41 - 1, -sqrt(41)/2)   # Area to left t1 in non-central t
```

```
[1] 0.7781911
```

Commands for finding the power at $n = 41$, 42, and 43 and the corresponding power values are provided below. Note that $n = 43$ is the smallest value of n to return a power (0.8011974) greater than 0.80.

```
pt(qt(0.01, 41 - 1), 41 - 1, -sqrt(41)/2)  # power with n = 41
```

```
[1] 0.7781911
```

```
pt(qt(0.01, 42 - 1), 42 - 1, -sqrt(42)/2)  # power with n = 42
```

```
[1] 0.7899524
```

```
pt(qt(0.01, 43 - 1), 43 - 1, -sqrt(43)/2)  # power with n = 43
```

```
[1] 0.8011974
```

After the students solve the problem using statistical software, have them verify their answer with the Shiny app at https://mathr.math.appstate.edu/shiny/Power2/. Students will specify the alternative hypothesis ($H_A : \mu < \mu_0$), select the appropriate test (t-test), enter the hypothesized mean (300), type the true mean (275), type the population standard deviation (50), select the significance level (0.01), and experiment with increasing or decreasing the sample size to find the minimum value of n to achieve a power greater than or equal to 0.80. A screen capture of the Shiny app with $n = 43$ is shown in Figure .

ADDITIONAL EXAMPLES

Once the students are comfortable using the Shiny app, they will be able to answer a variety of questions relating to power, sample size, the probability of making a type II error (β), and significance level (α) by experimenting with different inputs in the Shiny app. The following questions work well with the app.

a. Before a drug is administered, the lead physician states she believes the standard deviation for the patients administered the new drug will have a total cholesterol standard deviation between 40 mg/dL and 70 mg/dL. Use this new information to recompute your sample size requirements for the experiment using a t-test. Develop new recommendations and explain your new sample size requirements to meet the stated objectives.

Partial Answer: A conservative answer would use the total cholesterol standard deviation of 70 mg/dL to ensure the probability of detecting a reduction of 25 mg/dL is at least 80%. Using 70 mg/dL as the standard deviation, a sample size of $n = 82$ will detect a 25 mg/dL reduction in total cholesterol 80.33% of the time when the significance level is 0.01.

b. What sample size is need to ensure the type II error is no greater than the type I error?

Partial Answer: A sample size of $n \geq 173$ will return β (the probability of making a type II error) $< \alpha = 0.01$.

References

Arnholt, A. T. (2019). Using a shiny app to teach the concept of power. *Teaching Statistics*, *41*(3), 79–84. https://doi.org/https://doi.org/10.1111/test.12186

Ugarte, M. D., Militino, A. F., & Arnholt, A. T. (2015). *Probability and Statistics with R, Second Edition* (2 edition). Boca Raton: Chapman; Hall/CRC.