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

true

Last edit: December 16, 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 1

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.

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

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

```
(CriticalF <- qf(0.95, 2 - 1, 20 - 2))
```

```
[1] 4.413873
```

```
# Power for lambda
(pf(CriticalF, 2 - 1, 20 - 2, 11.25, lower = FALSE) -> POWER)
```

```
[1] 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.

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

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

```
(CriticalF <- qf(0.95, 2 - 1, 9 + 8 - 2))
```

```
[1] 4.543077
```

```
# Power for lambda
(pf(CriticalF, 2 - 1, 9 + 8 - 2, 9.5625, lower = FALSE) -> POWER)
```

```
[1] 0.8236915
```

The same calculations are performed next under the assumption that $\sigma = 30$ mg.

$$\begin{aligned}\lambda &= \frac{(\mu_1 - \mu_2)^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{\sigma^2} = \frac{SS_{\text{Hypothesis}}(\text{population})}{\sigma^2} \\ &= \frac{\sum_{1=1}^{a} n_i(\mu_{i\bullet} - \bar{\mu}_{\bullet\bullet})^2}{\sigma^2} = \frac{10(100-85)^2 + 10(70-85)^2}{30^2} = 5\end{aligned}$$

$$\text{Power}(\lambda = 5) = P(F^*_{1,18,\lambda=5} \geq f_{0.95,1,18} = 4.4138734) = 0.5620066$$

```
(CriticalF <- qf(0.95, 2 - 1, 20 - 2))
```

```
[1] 4.413873
```

```
# Power for lambda
(pf(CriticalF, 2 - 1, 20 - 2, 5, lower = FALSE) -> POWER)
```

```
[1] 0.5620066
```

Since we need to achieve a power of 0.80, we need to increase the values for each sample. Next, we consider using $n_1 = 17$ and $n_2 = 17$ when $\sigma = 30$ mg.

$$\lambda = \frac{(\mu_1 - \mu_2)^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{\sigma^2} = \frac{SS_{\text{Hypothesis}}(\text{population})}{\sigma^2}$$
$$= \frac{\sum_{1=1}^{a} n_i(\mu_{i\bullet} - \bar{\mu}_{\bullet\bullet})^2}{\sigma^2} = \frac{17(100-85)^2 + 17(70-85)^2}{30^2} = 8.5$$

$$\text{Power}(\lambda = 8.5) = P(F^*_{1,32,\lambda=8.5} \geq f_{0.95,1,32} = 4.1490974) = 0.8070367$$

```
(CriticalF <- qf(0.95, 2 - 1, 17 + 17 - 2))
```

```
[1] 4.149097
```

```
# Power for lambda
(pf(CriticalF, 2 - 1, 17 + 17 - 2, 8.5, lower = FALSE) -> POWER)
```

```
[1] 0.8070367
```

If the standard deviation for chick weight is $\sigma = 30$ mg, the experimenter needs to have 17 chicks assigned to each group to detect mean differences between groups greater than 80% percent of the time. If the standard deviation for chick weights is as small as $\sigma = 20$ mg, the experimenter can assign 9 chicks to the first group and 8 chicks to the second group and still detect mean differences greater than 80% of the time. To save space, we selected the values of n_1 and n_2 that solved the problem for the different values of σ. Next we show how to compute the non-centrality parameter λ using the notion of the $SS_{\text{Hypothesis}}(\text{population})$ using R, but which will work with any statistical software that computes the sum of squares.

```
# Hypothesized means
hypmeans <- c(100, 70)
# Number of treatments
a <- length(hypmeans)
n1 = 17
n2 = 17
# Total number of experimental units
N <- n1 + n2
# Degrees of freedom for error
dferror <- N - a
# Sigma value
sigma <- 30
# 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   7650    7650 3.673e+31 <2e-16 ***
Residuals   32      0       0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
[1] 7650
```

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

```
[1] 8.5
```

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

```
[1] 4.149097
```

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

```
[1] 0.8070367
```

Solution using the Shiny app

To find the required samples sizes to test $H_0 : \mu_{\text{testosterone}} = \mu_{\text{isoandrostenolone}}$ versus $\mu_{\text{testosterone}} \neq \mu_{\text{isoandrostenolone}}$ when $\sigma = 30$ mg, launch the Shiny app found at https://shinyapp and enter the values of 100 and 70 separated with a comma in the H_A box, 10 and 10 separated with a comma in the n_1, n_2, n_a box, the value of 20 in the σ box, and use the slider to select a significance level of 0.05 as shown in Figure 1. Change the values for n_1 and n_2 to be as small as possible with a power value of at least 0.80.

The red density in Figure 1 depicts a central F-distribution with $2 - 1$ and $20 - 2 = 18$ degrees of freedom. The blue density in Figure 1 is a non-central F-distribution with $2 - 1$ and $20 - 2 = 18$ degrees of freedom and non-centrality parameter ($\lambda = 11.25$). The purple shaded area in Figure 1 is the significance level and the sum of all blue and purple shaded areas is the power (0.887). Changing the values for n_1 and n_2 to either 8 and 9 or 9 and 8 results in a power of (0.8224) as shown in Figure 2.

Finally, have the students verify that the minimum samples sizes for n_1 and n_2 are both 17 when $\sigma = 30$ using the Shiny app as shown in Figure 3.

EXAMPLE 2

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) $< \alpha = 0.01$.

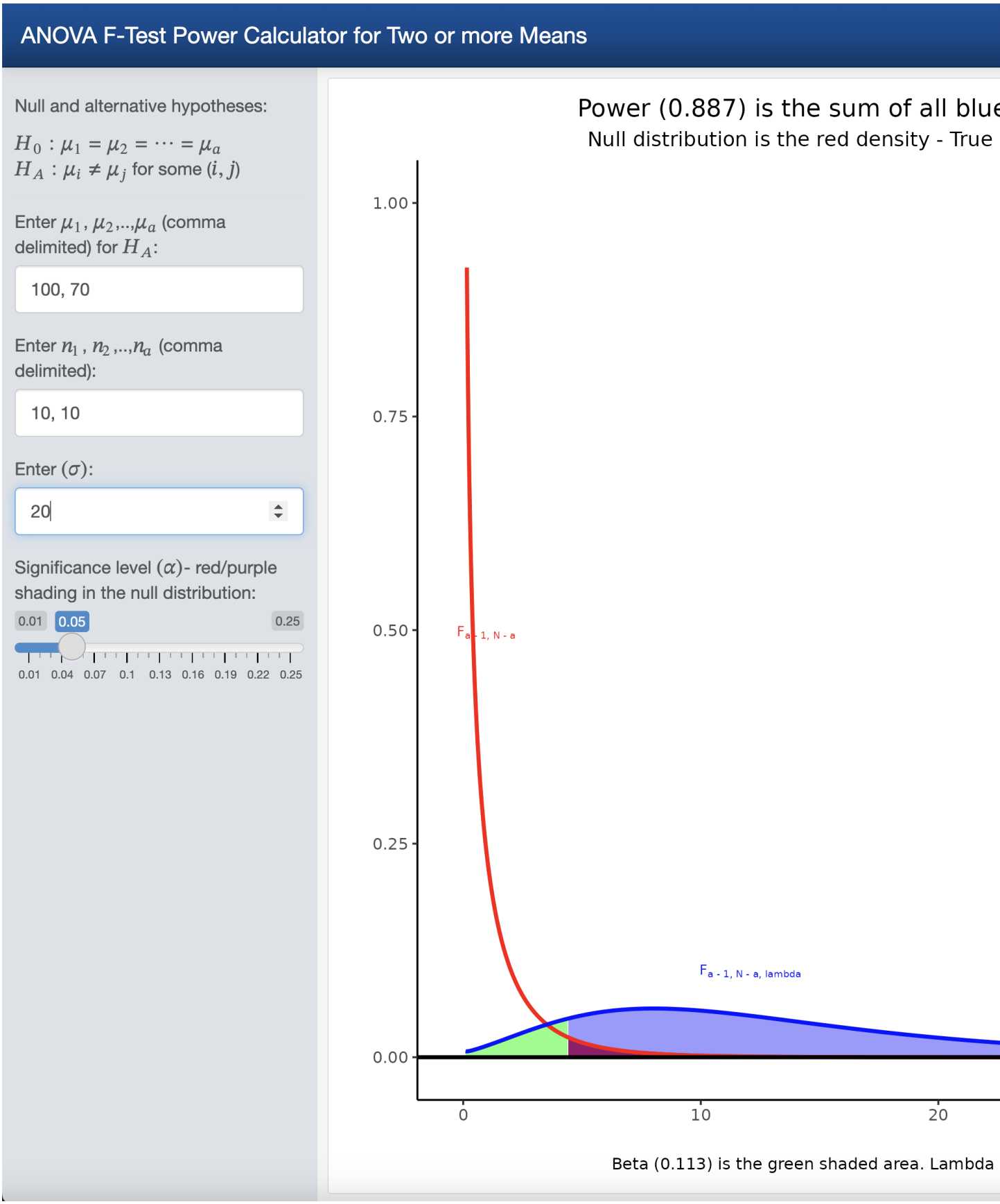

Figure 1: Power to detect specified means with given sample sizes and population standard deviation

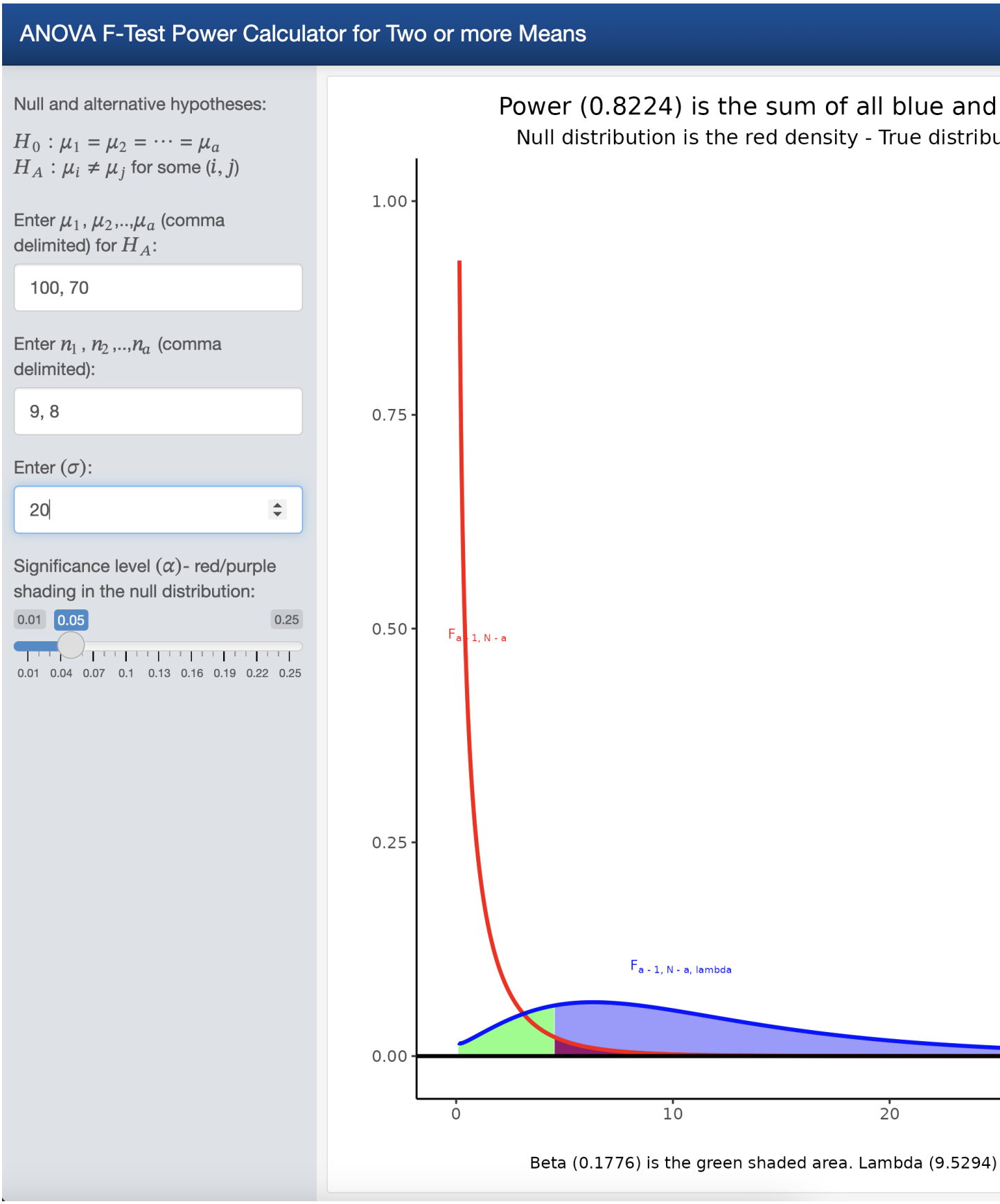

Figure 2: Power to detect specified means with given sample sizes and population standard deviation

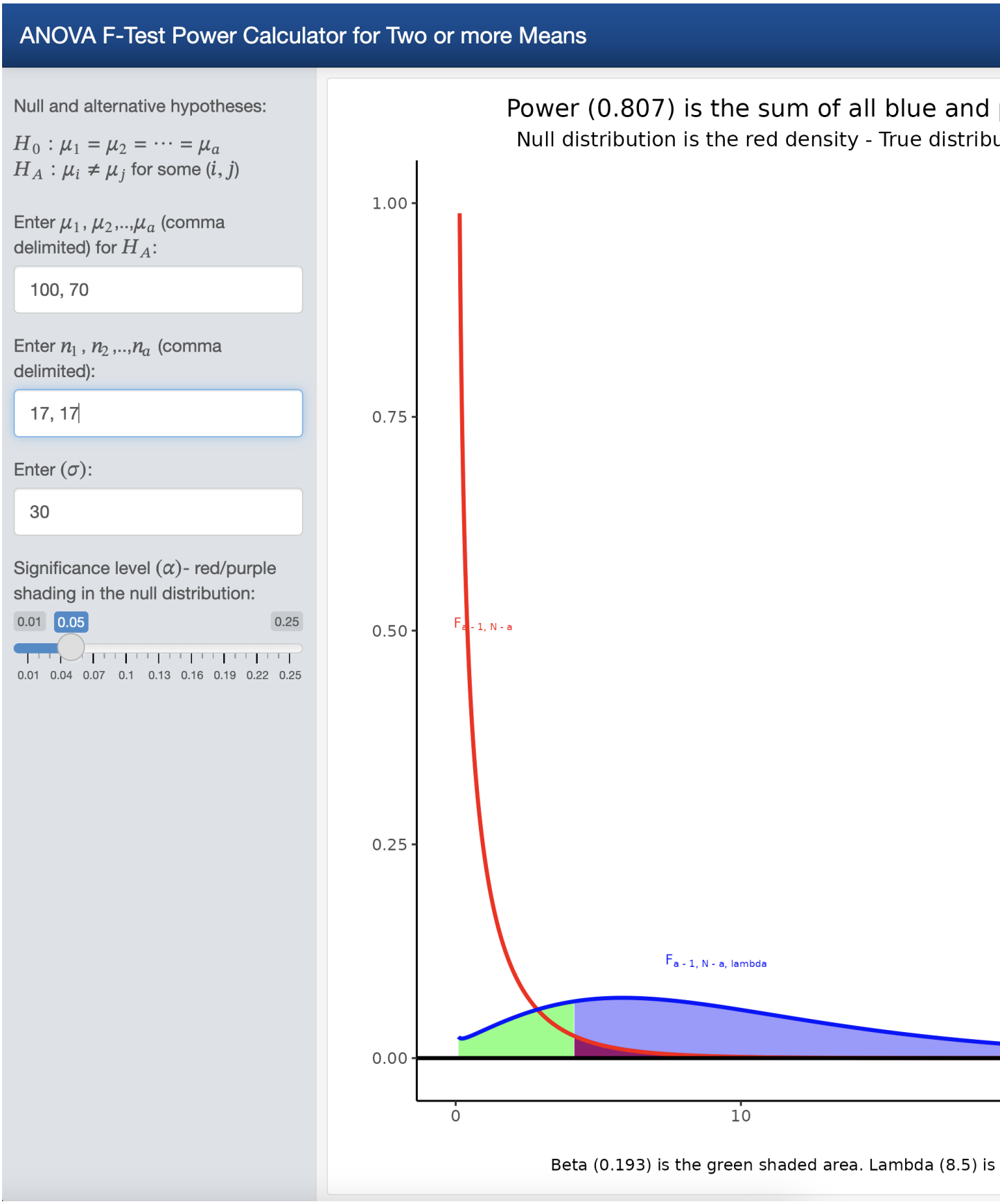

Figure 3: Power to detect specified means with given sample sizes and population standard deviation

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.