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

true

Last edit: February 23, 2025 at 01:41:03 PM

Abstract

Summary A Shiny app facilitates computing power and visualizing the non-central F distribution based on different population means, a single population standard deviation, different sample sizes, and any significance level. The Shiny app allows students to solve various scenarios by entering these parameters' values and by moving the significance slider.

Keywords: Power, Shiny app, significance level

INTRODUCTION

Following the GAISE guidelines (Carver et al., 2016), many instructors have added assessments in the form of small projects to the end of their courses to improve and evaluate student learning. It has been our observation that students will often select one of the last topics they have covered in their course for a project, often comparing many means when given the opportunity. When comparing many means, it is important the study allocate sufficient subjects/experimental units to each treatment to be able to detect differences in treatment means when those differences exist (power). The Shiny package (Chang et al., 2024) can be used to create web based applications. 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 non-centrality parameter (λ) is a measure of statistical difference between **population** means and how the F value computed in an ANOVA table is a measure of statistical difference between **sample** means. A Shiny app written by the authors 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. To aid with the connection between non-centrality parameters and test statistics, it is shown how the pooled variance t-test is a special case of the F-test when there are two treatments before generalizing to two or more 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; the τ_is are parameters unique to the i treatments called collectively treatment effects; and ϵ_{ij} are random errors associated with each observation. 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

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \quad \text{versus} \quad H_A : \mu_i \neq \mu_j \text{ for some } (i \neq j). \tag{2}$$

The notation that follows is adopted from Ugarte, Militino, & Arnholt (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 $\bar{Y}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij} = \sum_{i=1}^{a}\frac{Y_{i\bullet}}{N} = \sum_{i=1}^{a} n_{i\bullet}\bar{Y}_{i\bullet}/N$. For $a = 2$ treatments, the typical sum of squares for testing the hypotheses in (2) are shown to the left of the equivalence ($\equiv$) symbol in (5) where for a treatments $n_1 + n_2 + \cdot + n_a = N$, $df_{\text{Treatment}} = a - 1$, and $df_{\text{Error}} = N - a$. The representation on the right side of the $\equiv$ in (5) is the standard test statistic used to test the difference of two means when variances are assumed to be equal. To verify the equivalence, keep in mind that $\sum_{j=1}^{n_i} Y_{ij} = n_i\bar{Y}_{i\bullet}$ for $i = 1, 2$ and $(n_1 + n_2)\bar{Y}_{\bullet\bullet} = \sum_{i=1}^{2}\sum_{j=1}^{n_i} Y_{ij}$ and that $S_i^2 = \sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2/(n_1 - 1)$.

$$SS_{\text{Error}} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2 = \sum_{i=1}^{2}\left[\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2\right] = (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 \tag{3}$$

$$MS_{\text{Error}} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = S_p^2 \tag{4}$$

$$F = \frac{MS_{\text{Treatment}}}{MS_{\text{Error}}} = \frac{\dfrac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}{df_{\text{Treatment}}}}{\dfrac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2}{df_{\text{Error}}}} \equiv \frac{(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{S_p^2} \tag{5}$$

Rewriting the right side of (5) yields

$$F = \frac{(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})^2}{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \left[\frac{(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right]^2 = [t]^2. \tag{6}$$

The quantity in (5) measures the statistical differences between **sample** means. Let $SS_{\text{Hypothesis}}(\text{population})$ be 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 $\frac{n_1\mu_1 + n_2\mu_2}{n_1 + n_2}$. Then, replacing the sample means in (5) with the population means and replacing S_p^2 with σ^2 yields the non-centrality parameter of the F distribution,

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

The formula to find λ for testing equality of a means is

$$\lambda = \frac{SS_{\text{Hypothesis}}(\text{population})}{\sigma^2} = \frac{\sum_{i=1}^{a} n_i(\bar{\mu}_{i\bullet} - \bar{\mu}_{\bullet\bullet})^2}{\sigma^2},$$

which one obtains by replacing the sample means in the standard sum of squares treatments formula with corresponding population means. 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.

EXAMPLE 1

The following scenario is suitable for students:

Consider a fictitious experiment where one of two hormones (testosterone or isoandrostenolone) is administered for fifteen days to male chicks starting when the chicks are one day old. At the end of the fifteen days, the

experimenter hypothesizes the average weight of chicks that receive testosterone will be 100 mg and that 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

Although R (R Core Team, 2024) is used as the computational engine in what follows, most modern software packages can handle what is shown in the code. Start by asking the students to specify the null and alternative hypotheses they will use to test the mean difference in chicks' weights that were given the two hormones. The hypotheses are:

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

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 non-centrality 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. Although power is maximized when sample sizes are equal, in this example, $n_1 = n_2 = 9$ returned a power of 0.8476, while $n_1 = n_2 = 8$ returned a power of 0.7965. Consequently, we chose to use unequal sample sizes. It does make a difference whether the first treatment group or the second treatment group receives the smaller number of chicks. If the second group receives the smaller number of chicks, the resulting power would be 0.8223981. Since the groups are of different sizes, one must weight the means appropriately. That is $\bar{\mu}_{\bullet\bullet} = \frac{n_1\mu_1 + n_2\mu_2}{n_1 + n2} = \frac{9\times100 + 8\times70}{9+8} = 85.88235$.

$$\begin{aligned}\lambda &= \frac{(\mu_1 - \mu_2)^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{\sigma^2} = \frac{SS_{\text{Hypothesis}}(\text{population})}{\sigma^2} \\ &= \frac{\sum_{1=1}^{a} n_i(\mu_{i\bullet} - \bar{\mu}_{\bullet\bullet})^2}{\sigma^2} = \frac{9(100-85.88235)^2 + 8(70-85.88235)^2}{20^2} = 9.529412\end{aligned}$$

$$\text{Power}(\lambda = 9.529412) = P(F^*_{1,15,\lambda=9.529412} \geq f_{0.95,1,15} = 4.5430772) = 0.8223981$$

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

```
[1] 4.543077
```

```
# Power for lambda
(pf(CriticalF, 2 - 1, 9 + 8 - 2, 9.529412, lower = FALSE) -> POWER)
```

```
[1] 0.8223981
```

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

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

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

```
(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 more 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.

```r
# 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
```

```r
# 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 $H_A : \mu_{\text{testosterone}} \neq \mu_{\text{isoandrostenolone}}$ when $\sigma = 30$ mg, launch the Shiny app found at https://hasthika.shinyapps.io/anovaShinyApp/ 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 $n_1, n_2, \ldots, n_a$ 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. Additionally, the $P(\text{Type II error}) = \beta$ is depicted as the green shaded area in 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 which will appear in the title of the Shiny app in parentheses.

The red density in Figure 1 depicts a central F-distribution with $2 - 1 = 1$ and $20 - 2 = 18$ degrees of freedom. The blue density in Figure 1 is a non-central F-distribution with $2 - 1 = 1$ and $20 - 2 = 18$ degrees

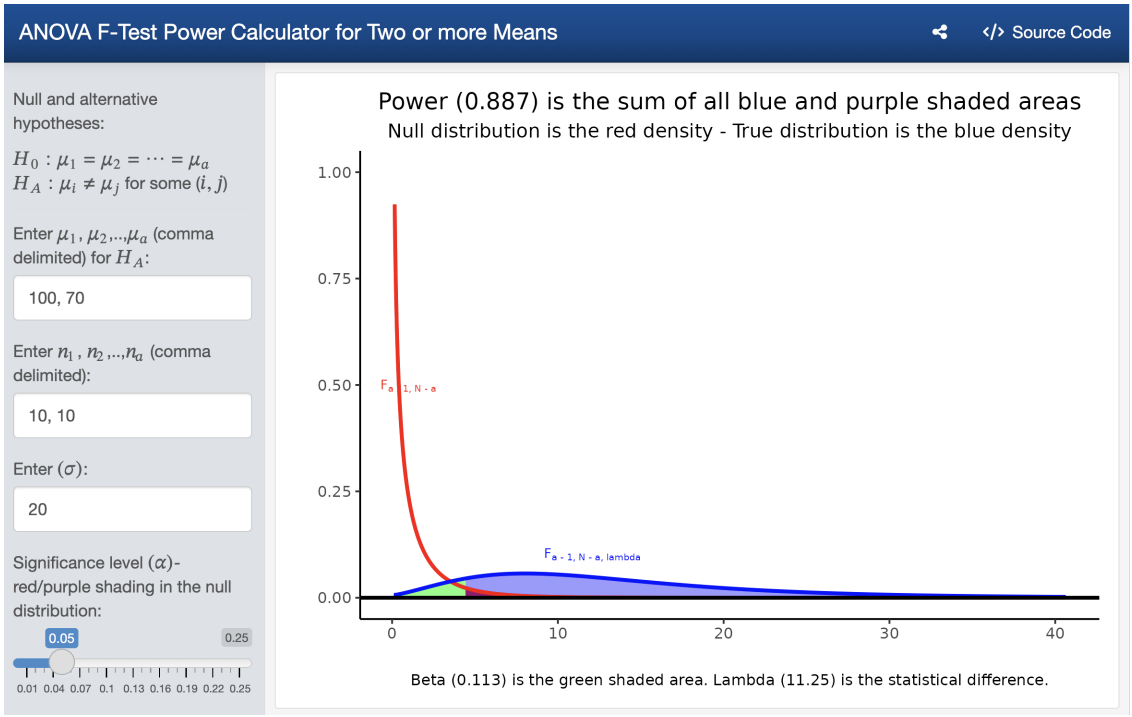

Figure 1: Power to detect a specified difference in population means with given sample sizes and population standard deviation

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

An educational researcher is interested in testing different tools to help his students master statistical concepts. The researcher hypothesizes there will be increases in mean performance for students on a standardized test from lecture alone (70), to lecture with statistical software (75), to lecture with statistical software and videos (80), to lecture with statistical software, videos, and shiny apps of (85). If the population standard deviation for the researcher's students on the standardized test is 15, find the minimum number of students required for each group to be able to have a power value of at least 0.80.

Figure 4 shows 70, 75, 80, and 85 entered for μ_1, μ_2, μ_3, and μ_4 in the H_A box, the value of 15 in the σ box, and sliding the significance level to 0.05. Power is maximized when all treatment groups have the same number of experimental units. Consequently, if there are no restrictions on the allocation of resources, the experimenter should allocate an equal number of experimental units to each of the a treatments. Then, experiment with different values for the sample sizes for each treatment until the power is greater than or equal to 0.80. Our trial and error method started with $n_1 = n_2 = n_3 = n_4 = 10$ which yielded too small a power. As a guess, we doubled the values to $n_1 = n_2 = n_3 = n_4 = 20$ and noted we were just a little short of our desired power value. Next we tried $n_1 = n_2 = n_3 = n_4 = 25$ and observed this was over the target value of 0.80. We continued experimenting lowering the values until we arrived at $n_1 = n_2 = n_3 = n_4 = 21$ which returned a power of 0.8082.

Answer: $n_1 = n_2 = n_3 = n_4 = 21$.

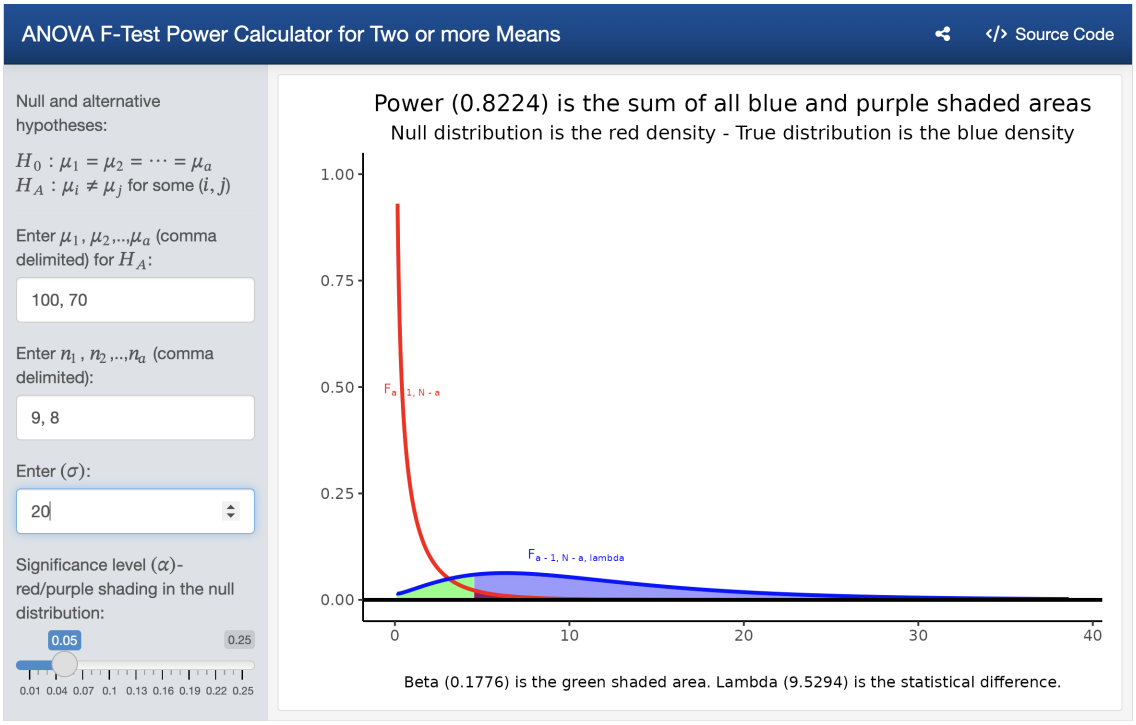

Figure 2: Power to detect a specified difference in population means with given sample sizes and population standard deviation

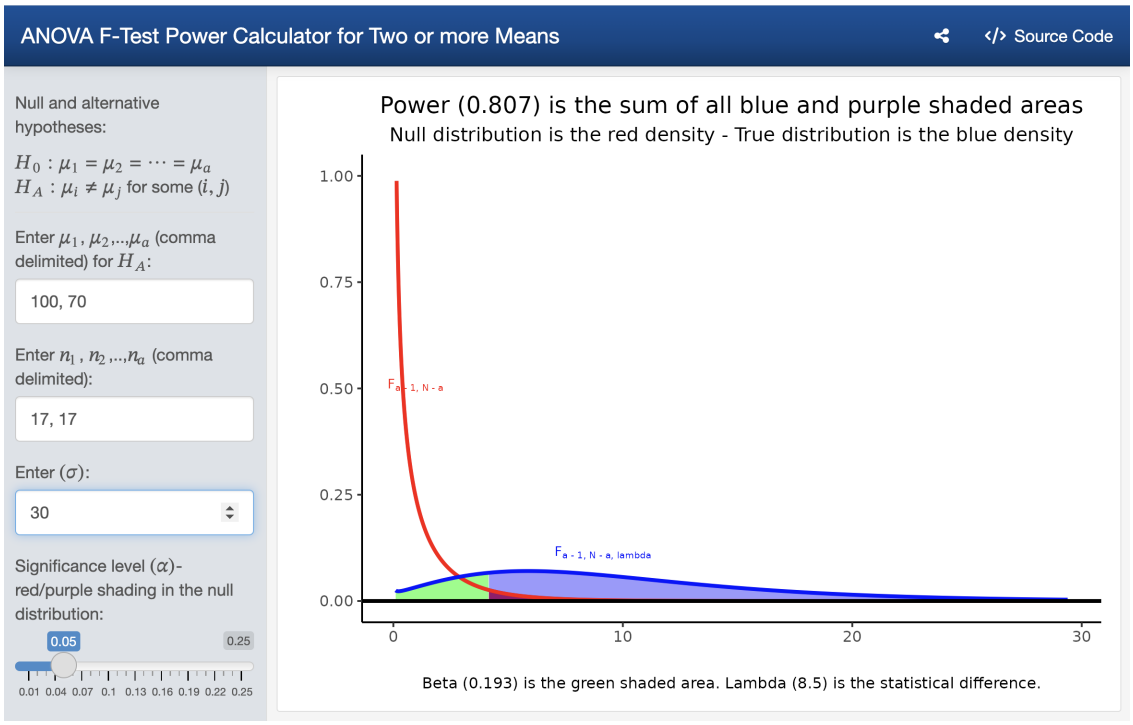

Figure 3: Power to detect a specified difference in population means with given sample sizes and population standard deviation

Table 1: Process of trial and error for desired power

n	Power
10	0.4396
20	0.7856
25	0.8797
22	0.8287
21	0.8082

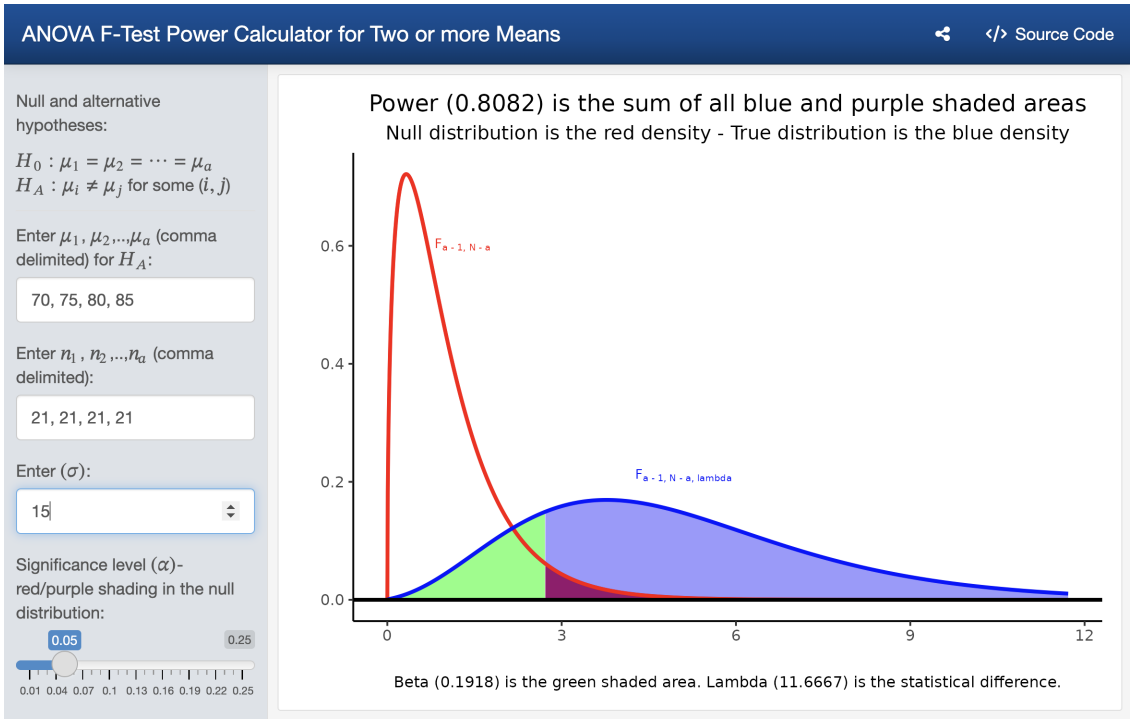

Figure 4: Power to detect a specified difference in population means with given sample sizes and population standard deviation

We note in passing that computing power for full-rank general linear models is also possible using the same paradigm where $\lambda = \frac{SS_{\text{Hypothesis}}(\text{population})}{\sigma^2}$. For details see section 12.10 of Ugarte et al. (2015).

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

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., ... Wood, B. (2016). Guidelines for assessment and instruction in statistics education (GAISE) college report 2016.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., ... Borges, B. (2024). *Shiny: Web application framework for r*. Retrieved from https://shiny.posit.co/

R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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