Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 18, August 2022 ISSN 1531-7714

Practical t-test Power Analysis with R

Teck Kiang Tan, *National University of Singapore*

Power analysis based on the analytical t-test is an important aspect of a research study to determine the sample size required to detect the effect for the comparison of two means. The current paper presents a reader-friendly procedure for carrying out the t-test power analysis using the various R add-on packages. While there is a growing of R users in the academic that uses R as the base for carrying out research, there is a lack of reference that discusses both frequentist and Bayesian approaches and point out their distinct features for t-test power analysis. The practical aspects of the consequences of unequal variances and sample sizes are often neglected and this paper discusses and illustrates using the graphical power curve. A written R function and several programs are used to illustrate the usefulness of choosing an appropriate sample size under the frequentist approach. The Bayes factor is introduced to show its expediency to generate the required sample size under the Bayesian approach. Researchers and practitioners with intervention research using the t-test to carry out hypothesis testing will find this paper a commendable power analysis reference to design their project.

Keywords: Power Analysis, t-test, Package R, Bayesian Approach, Frequentist Approach

Introduction

Power analysis is often referred to as the process of determining the sample size for a study (Connelly, 2008). Selecting an appropriate sample size is a crucial step in designing a successful study, for instance, for an intervention study that requires comparing the difference in two means to examine the magnitude of the effect where the main objective is to compare the pre-and post-test mean. Two-group comparison has been and is always at the heart of many researchers. The two-sample t-test is arguably the most commonly used statistical test in research for carrying out the comparison of two groups, two occasions, two cohorts, or two-time points to examine carry-over, period, and cohort effect. Nuijten et al. (2016) reported the meta-analysis results after examining 258,105 p-values of the psychology journals between 1985 and 2013, out of which 26% belonged to a t-statistic. The way t-test power analysis is carried out becomes crucial for a study to determine the appropriate sample size. A study to compare the group of two means that provides insufficient sample sizes may not have sufficient statistical power to detect meaningful effects and produce unreliable results to answer the research question of whether the result of the intervention is a “true” one. On the other hand, a study with an excessive sample size wastes resources. Choosing the “right” sample size increases the chance of detecting an effect, and at the same time ensures that the study is cost-effective (Carneiro, 2003; Legg & Nagy, 2006). Therefore, it is a good practice to perform a t-test power analysis as earlier as during the study design stage.

Power analysis for the t-test is not a straightforward task simply using a formula stated in the standard textbook to calculate the required sample sizes for the two groups under study (e.g. Desu & Raghavarao,1990) as many factors could affect the results of choosing the appropriate sample size. Some factors are more sensitive for one study and lesser for another. While sample size calculation is usually carried out using software, software that could carry out graphing and at the same time provide the functions to vary the relevant factors to examining sensitivity analysis is always preferred. While there is an exponential growth of R users (Datanami, 2020; Stack Overflow Blog, 2017;) using this freeware R package with the development of the R add-on packages that perform power analysis (Champel, 2020; Dong & Maynrd, 2013; Fu, 2021; Kohl, 2020; Shen, 2022; Zhong & Mai,2021) and providing the graphical features (Sarkar, 2008; Wickham, 2016), this software turns out to be an ideal one.

The main focus of the paper is to call attention to the practical guidelines for carrying out power analysis for the analytical t-test using the various add-on R software. The t-test assesses the statistical significance of a specific value or the difference between two independent population means or the difference between means of matched pairs. From the frequentist viewpoint, the technical definition of power is that it is the probability of detecting a "true" effect when it exists. Put another way, the power of a hypothesis test is defined as the probability that the t-test will reject the null hypothesis of equality of two means, assuming that the null hypothesis is false. That is, if an effect is real, what is the probability that analysis will judge that the effect is statistically significant. Two main approaches to attaining a required sample size, the frequentist, and the Bayesian approach are discussed to determine whether there is an effect when comparing two means.

Frequentist Approach

Classical methods for sample size determination of the t-test when testing hypotheses are typically related to the use of the power function. The goal is to determine the minimal size of a sample such that, for a fixed Type I error probability, the chance of correctly rejecting the null hypothesis is sufficiently large. From the pure statistical requirements and qualifications, the traditional frequentist method of estimating sample sizes require the specification of the following four quantities (Livingston & Cassidy, 2005):

(1) Guess or Predict Effect Size

Knowledge about effect size is one crucial piece of information to carry out power analysis. It is a quantitative measure of the magnitude of the study effect. The standardized effect size Cohen's d is δ/σ, where $\delta = |\mu - \mu_0|$ is the treatment difference and δ is the standard deviation. When the effect size is unknown, the description of the magnitude of the Cohen's d effect size (Cohen, 1988) to quantify the level of effect sizes into the small, medium, and large is often used to set the range of the effect size (Refer to Appendix E for the range of Cohen's effect size).

(2) Tolerance Type I Error / Alpha (α) / Significance level

Significance level (α) refers to the probability of falsely rejecting the null hypothesis even though it is true. That is the probability of a Type I error (false positive). The lower the significance level, the more likely it is to avoid a false positive and the more samples needed. The standard setting for α is 0.05.

(3) Desired Power (1-β)

Power is the probability of correctly rejecting the null hypothesis if it is false. That is the probability of detecting a true difference when it exists. Power = 1 - β, where β is the probability of a Type II error (false negative). The higher the power, the more likely it is to detect an effect if it is present, and larger sample size is needed. The standard power setting is normally set to 0.80.

(4) One-Sided or Two-Sided Test

For null hypothesis $H_0\colon \mu = \mu_0$ and the alternative hypothesis $H_1\colon \mu \neq \mu_0$, it is a two-sided test. When the null hypotheses become either $H_0\colon \mu < \mu_0$ or $H_0\colon \mu > \mu_0$, it is a one-sided test. Whether using one-sided or two-sided depending on the research concern.

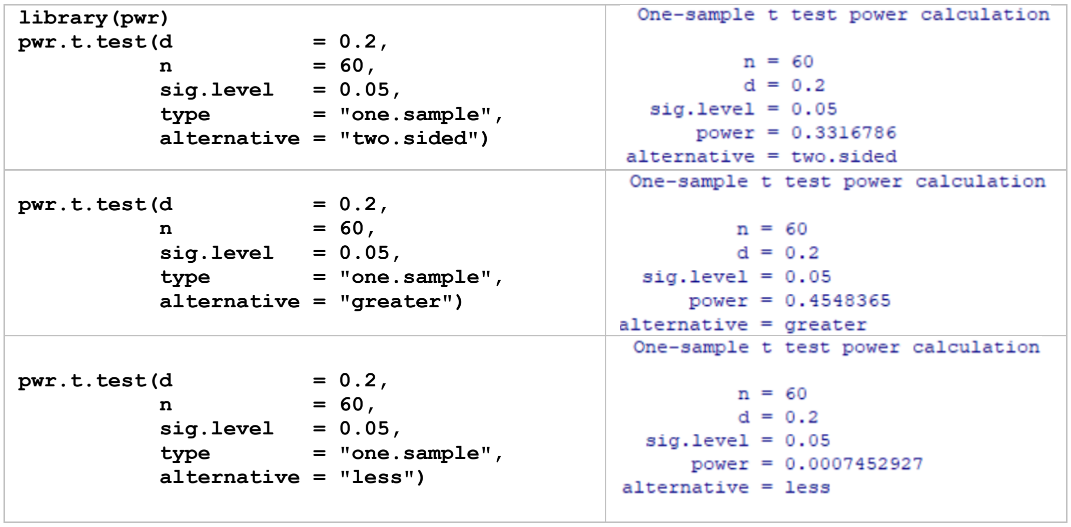
Power Analysis of Single Sample – t-test

Suppose an education researcher is interested in examining whether using a pedagogical approach improves mathematical ability to attain an ability level with an overall mean of μ_0. The study would like to establish whether there is enough power to detect this attained level by setting $H_0\colon \mu = \mu_0$ and the alternative hypothesis $H_1\colon \mu \neq \mu_0$. For one-sided testing, the null hypotheses become either $H_0\colon \mu < \mu_0$ or $H_0\colon \mu > \mu_0$. To determine the necessary sample size for the hypothesis, the researcher has to specify the maximum acceptable risk of rejecting $H_0$when it is true, α, as well as the maximum acceptable risk of failing to reject H_0 when it is false, β and the effect size, d.

There are at least four R packages to carry out the power analysis using the frequentist approach. The function power.t.test in R base (R Core Team, 2021), the function pwr.t.test from the package pwr (Champely, 2020), the function wp.t from the package WebPower (Zhang & Mai, 2021; Zhang & Yuan, 2018) and the function pwr.welch.t.test from the MKpower package (Kohl, 2020). For illustration purposes, this paper concentrates on using the function pwr.t.test for one-sample and two-sample t-test, and function pwr.welch.t.test for Welch's t-test, the appendices give the syntaxes and outputs for other functions.

The function pwr.t.test consists of 6 arguments. The argument n represents the sample size, the argument d represents the Cohen's d effect size, the argument sig.level and the argument power specifies the type I error probability (α) and one minus II error probability (i.e. power) respectively. The type argument stipulates whether it is a one-sample or two-sample t-test and the argument alternative specifies whether it is a "two.sided" (default), "greater" or "less" one-sided test.

The following three examples generate the values of power for one-sample hypothesis testing for two-sided and one-sided tests. Cohen's d is specified as 0.2 (d=0.2), the significant level is 0.05 (sig.level=0.05) and the sample size is 60 (n=60). For a two-sided t-test, the argument alternative is specified as "two-sided", for one-sided, the arguments are specified as either "greater" or "less". While there are six basic arguments for the pwr.t.test function, the omitted argument power automatically calculate the value of the power for the specified settings. The values of power or the two-sided, one-sided greater, and less are 0.33, 0.45, and 0.0007 respectively.



```
library(pwr)
pwr.t.test(d           = 0.2,
           n           = 60,
           sig.level   = 0.05,
           type        = "one.sample",
           alternative = "two.sided")
```

```
     One-sample t test power calculation

              n = 60
              d = 0.2
      sig.level = 0.05
          power = 0.3316786
    alternative = two.sided
```

```
pwr.t.test(d           = 0.2,
           n           = 60,
           sig.level   = 0.05,
           type        = "one.sample",
           alternative = "greater")
```

```
     One-sample t test power calculation

              n = 60
              d = 0.2
      sig.level = 0.05
          power = 0.4548365
    alternative = greater
```

```
pwr.t.test(d           = 0.2,
           n           = 60,
           sig.level   = 0.05,
           type        = "one.sample",
           alternative = "less")
```

```
     One-sample t test power calculation

              n = 60
              d = 0.2
      sig.level = 0.05
          power = 0.0007452927
    alternative = less
```

Power Analysis Interpretation – Power Curve

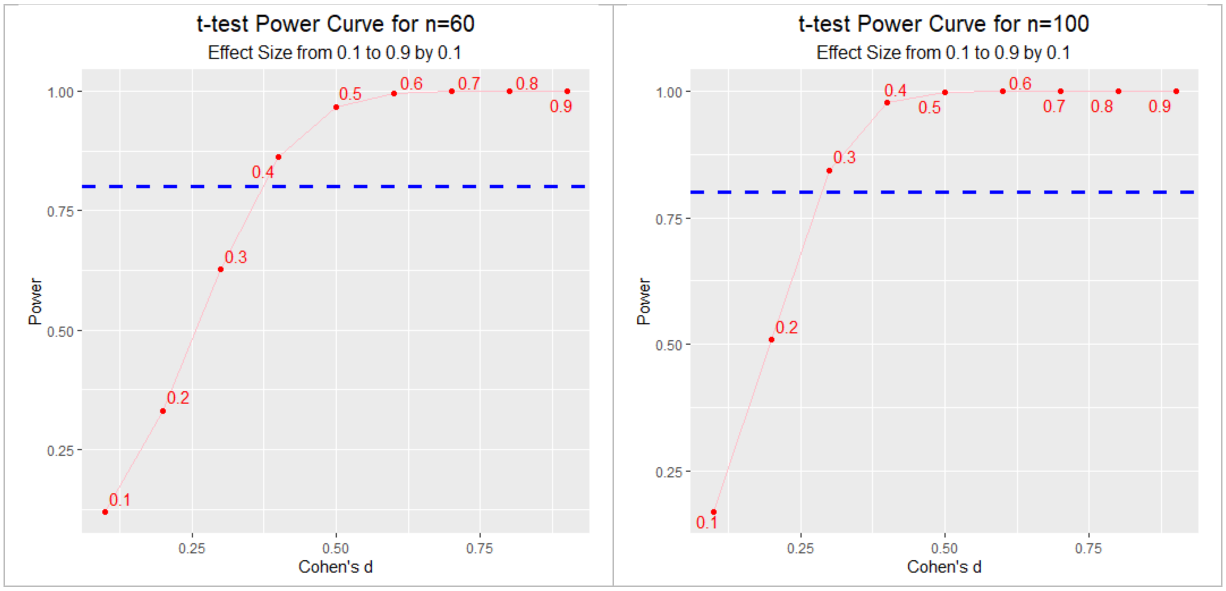
The above section generates three power analysis outcomes using the pwr.t.test function. This procedure can turn into a tedious process if the aim is to examine a series of power analysis outcomes by varying a factor, say the effect size, to examine the changes in the value of power. Most often, the power curve is a good choice to examine the changes in effect size and sample size that impact the power of a study. While the power curve is a useful line plot graphical representation to assess and determine an appropriate sample size or the power for a study, it is not readily available for the pwr package. As such, an R function named "PowerCurve.OneSample.FixedN" is written to graph the power curve for the one-sample t-test. This function uses three R packages, pwr, ggplot2, and ggrepel to produce an attractive colored power curve. The syntax of this function is listed in Appendix A, A1. The default setting of effect size for this function varies from 0.1 to 0.9 with an increment of 0.1. The function pwr::pwr.t.test calculates the power. The various geom functions from the ggplot2 package produce the power curve, and the package ggrepel makes use of the function geom_text_repel to ensure the effect sizes in numeric numbers are printed in red color in the power curve graph do not overlap. This function takes in a numeric value input that represents the expected sample size. The following syntax generates two power curves with the specification of 60 and 100 sample sizes.

```
n <- 60
PowerCurve.OneSample.FixedN(n)
n <- 100
PowerCurve.OneSample.FixedN(n)
```

The power curve with the specification of the sample size n as 60 and 100 produces the left-hand and right-hand of Figure 1 respectively. The x-axis represents Cohen's d and the y-axis represents the power. The dotted blue horizontal line positioned at 0.80 of the y-axis represents the general cutoff value of power. The red printed numeric values represent the effect size that runs from .1 to .9. The corresponding red dots represent the effect size to show if the dot lies above the blue line indicating its power is higher than the cutoff power of 0.8. and below it, is lower than the cutoff .8. Both the power curve shows that the higher the effect size, the higher the power. Comparison of both power curves, the power curve for a sample size of 60 shows that given the effect size value of .4, the power is above .8. However, for a higher sample size of 100, the effect size reduces to .3. This is a general expectation of power analysis that the higher the expected effect size for a study, the requirement for sample size is lower. The underlying reason is straightforward. If a study expects to have a large effect, it does not need a large sample size to verify it. On the contrary, if a researcher is unsure of the effect or the effect is expected to be small, it is good to get a larger sample size. The sensitivity of power is also demonstrated by the comparison of the two power curves. For a given sample size of 60, the power becomes flat with little changes starting at the effect size at 0.5, whereas, for a larger sample size of 100, the upper asymptotic level happens earlier at 0.4. The practical implication is that a study with an expected effect size of .4, say based on a past study, will tend to choose a sample size of 60. If past studies indicated effect size generally lower than .4, a researcher will tend to choose a higher sample size of 100. This is to ensure a larger sample is more likely and safer to ensure whether there is an effect.

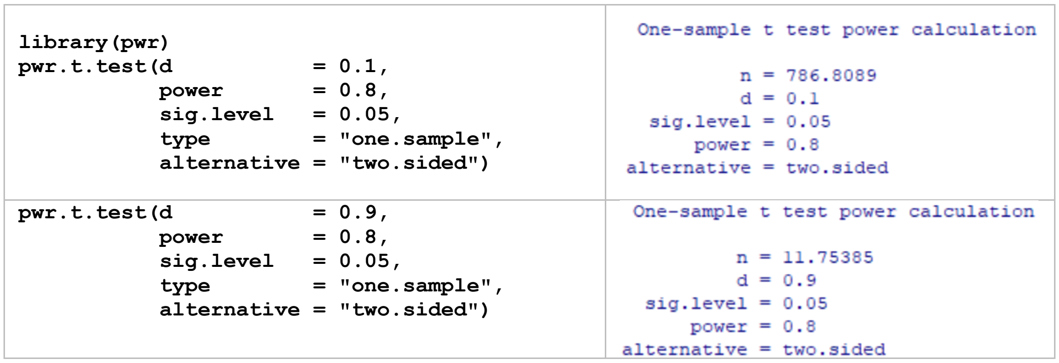
For a study with a cost constraint restriction on the upper limit of sample size, the above power analysis is helpful to give information with regards to the required power and expected effect size. While a researcher faces the situation of not knowing the effect size as there is no literature to base on to indicate the level of effect size, the concern becomes what is the range of sample size the researcher can rely on by varying the effect size with a fixed power to ascertain the required sample size. The following two examples use the same function

Figure 1. Power Curves for One-Sample t-test with sample sizes set to 60 and 100



pwr.t.test by omitting the argument n, instead of the argument power, specifying power to 0.8, stating the Cohen's effect size to 0.1 and 0.9, producing the required sample size of 787 and 12 respectively. This wide difference in sample size that is due to varying the effect size shows the importance of effect size in determining the sample size.

A graphical output showing the desired sample with a given power would be helpful to further extend the understanding of the previous two examples by specifying a range of effect sizes. The R program to generate the power graph is provided in Appendix A, A2. Specifying power as .6 and .8 produces the right and left graphs of Figure 2 respectively. In comparison to the setting of power at .6, and .8, the results of the two graphs indicate the higher the required sample size, the higher the power. These two power curves also indicate the higher the expected effect size, the lower the required sample size, and vice versa. The sensitivity in the required sample size is also indicated in these two graphs. Setting power at .6, the largest difference in the required sample size of 109 (198-89) in the effect size lies between .2 to .3. For power fixed at .8, a slightly smaller largest difference in the sample size is 68 (124-56). The power analyses point out that the major factor that can affect sample size is the effect size. In particular, for a move of effect size in the lower range, the impact on sample size is greater. These two graphs confirm the most sensitive area of effect size that causes a large change in the required sample size is between .2 to .4. If the size of the effect is known to be around say between .7 and .8, a researcher will be more comfortable as the changes in sample size are much easier to control. On the contrary, if the effect size of past studies shows lower effect sizes around .2, the impact on the required sample size can turn out as a major issue as the decision on the required sample size can vary tremendously with a small change in effect size.



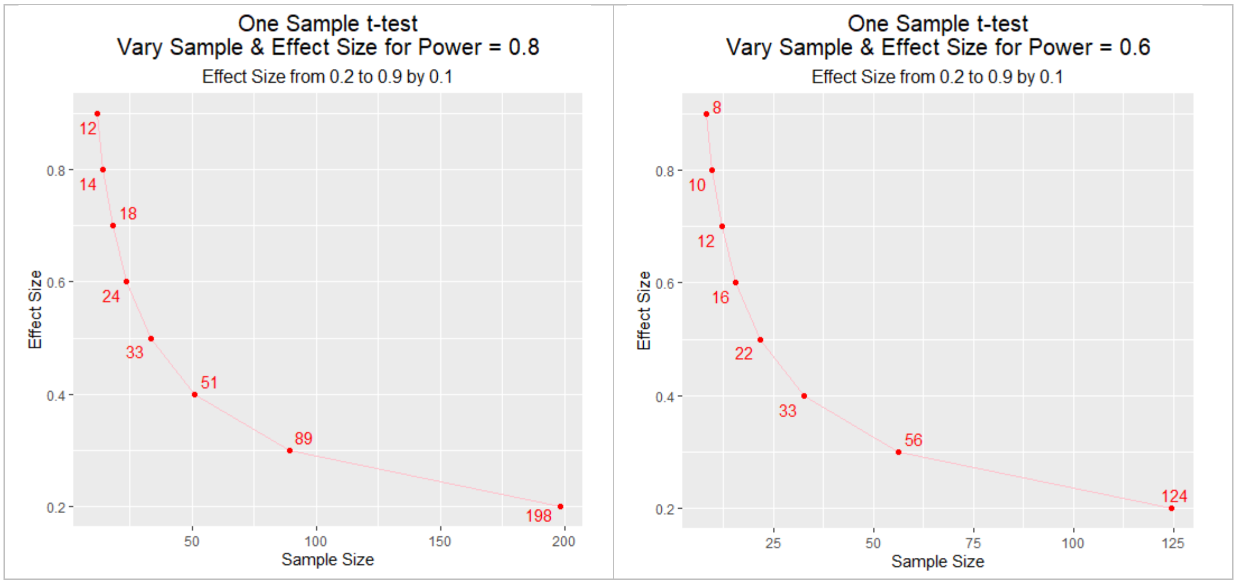
```
library(pwr)
pwr.t.test(d           = 0.1,
           power       = 0.8,
           sig.level   = 0.05,
           type        = "one.sample",
           alternative = "two.sided")
```

```
     One-sample t test power calculation

              n = 786.8089
              d = 0.1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

```
pwr.t.test(d           = 0.9,
           power       = 0.8,
           sig.level   = 0.05,
           type        = "one.sample",
           alternative = "two.sided")
```

```
     One-sample t test power calculation

              n = 11.75385
              d = 0.9
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

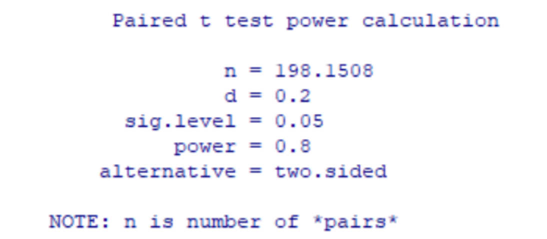
Figure 2. Varying Effect Size with A Fixed Power to Generate Sample Size



Two Sample Paired and Independent t-test

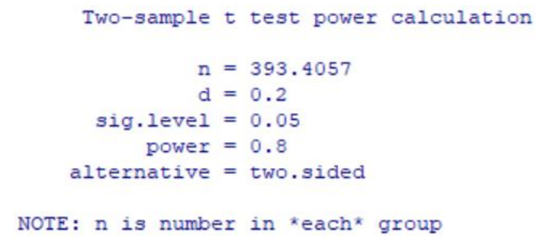
A paired t-test aims to examine whether there is a difference between two means for the same group of subjects. Often the two means are separated by time, occasion, or condition. For instance, a comparison of pre-intervention and post-intervention to examine whether there is a shift in mean after the intervention. The independent two-sample t-test, on the other hand, is to determine whether there is a statistically significant difference in means between two unrelated groups. Calculate the required total sample size for paired t-test and independent t-test is to specify type=”paired” for the former and the latter type=”two.sample” under the pwr.t.test function. The required sample sizes for the paired t-test and independent t-test are 198 and 393 respectively for the same specifications stated in the one-sample t-test power analysis. As paired t-test using the same subjects for a study, the required sample size is about half of that under an independent t-test. This is understandable as a within-subject study generally requires a lesser sample size.

```
library(pwr)
pwr.t.test(d=0.2,power=.8,sig.level=.05,
           type="paired",
           alternative="two.sided")
```



```
     Paired t test power calculation

              n = 198.1508
              d = 0.2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*
```

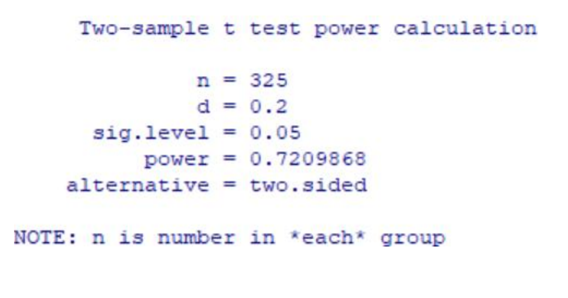
```
library(pwr)
pwr.t.test(d=0.2,power=.8,sig.level=.05,
           type="two.sample",
           alternative="two.sided")
```



```
     Two-sample t test power calculation

              n = 393.4057
              d = 0.2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

To calculate the power given the sample size for an independent t-test, omitted the argument power will generates the power value. The following example generates a power of .72 given a sample size of 325 and effect size of .2.

```
library(pwr)
pwr.t.test(d=0.2,n=325,sig.level=.05,
           type="two.sample",
           alternative="two.sided")
```



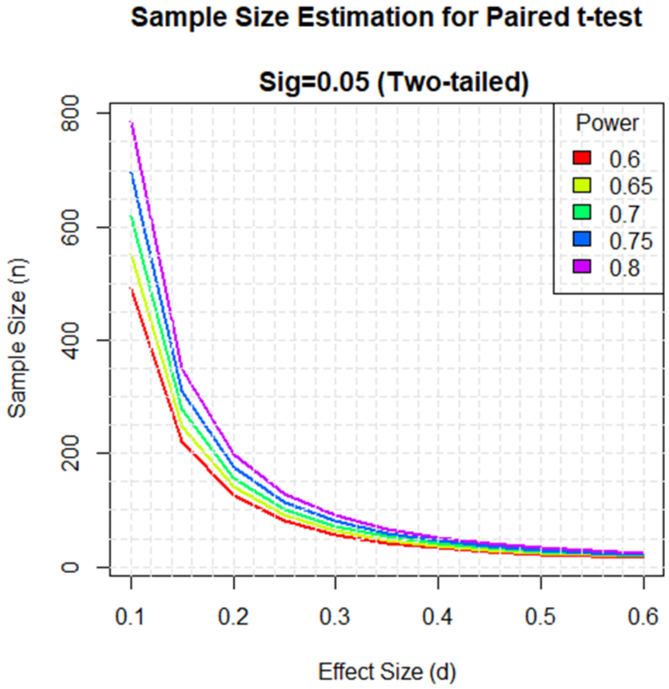
```
     Two-sample t test power calculation

              n = 325
              d = 0.2
      sig.level = 0.05
          power = 0.7209868
    alternative = two.sided

NOTE: n is number in *each* group
```

Figure 3 shows the effect on sample size by varying the effect size and power for the paired t-test. The syntax for generating the power curve and the table of sample size is listed in Appendix C. The power curve shows the level of effect size primarily determines the size of the sample required. An effect size of .1 with a power of .8 generates a sample size of 787 whereas a higher effect size of .6 with the same power of .8 reduces the required sample size to 24. Similar exponential decreases in sample size are noted for all the ranges of power from .6 to .8. The difference in sample size is wider for small effect size is also observed.

Practical research studies generally face unequal sample sizes, this can happen earlier at the planning stage. Sample determination for unequal sample size could be determined using the function pwr.t2n.test under the package pwr. The following shows two examples of calculating the value of power under the unbalanced independent t-test. The first example specifies the two sample sizes for n1=350 & n2=300 and n1=400 & n2=250 for the second example. While the total sample size for both examples is 650, the first example generates a higher power of .72 and the second example produces a lower power of .69.

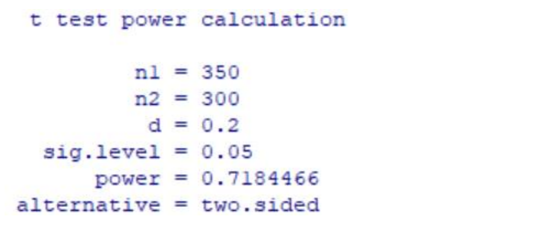
The above results indicate that when the ratio of n1 and n2 is closer to one for the first example (350/300 = 1.17), the power is higher at .72, and when the ratio is further away from equality (400/250 = 1.6) for the second example, the power becomes lower at .70. The concern becomes what are the changes in power for a study with a high and a low effect size. Figure 4 prints three graphs setting effect sizes at 0.2, 0.3, and 0.4. The simulation of the ratio of sample size varies from 1:5 to 5:1, with a total

Figure 3. Power Curve: Paired t-test Sample Size Estimation



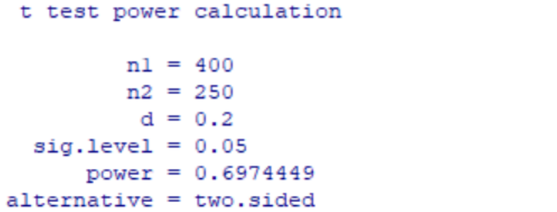
Effect	Power				
Size	0.60	0.65	0.70	0.75	0.80
0.10	492	552	620	696	787
0.15	220	247	277	311	351
0.20	125	140	157	176	199
0.25	81	90	101	113	128
0.30	57	64	71	80	90
0.35	42	47	53	59	67
0.40	33	37	41	46	52
0.45	27	30	33	37	41
0.50	22	24	27	30	34
0.55	19	21	23	25	28
0.60	16	18	20	22	24

```
library(pwr)
pwr.t2n.test(d=0.2,n1=350,n2=300,
             alternative="two.sided")
```



```
 t test power calculation 

          n1 = 350
          n2 = 300
           d = 0.2
   sig.level = 0.05
       power = 0.7184466
 alternative = two.sided
```

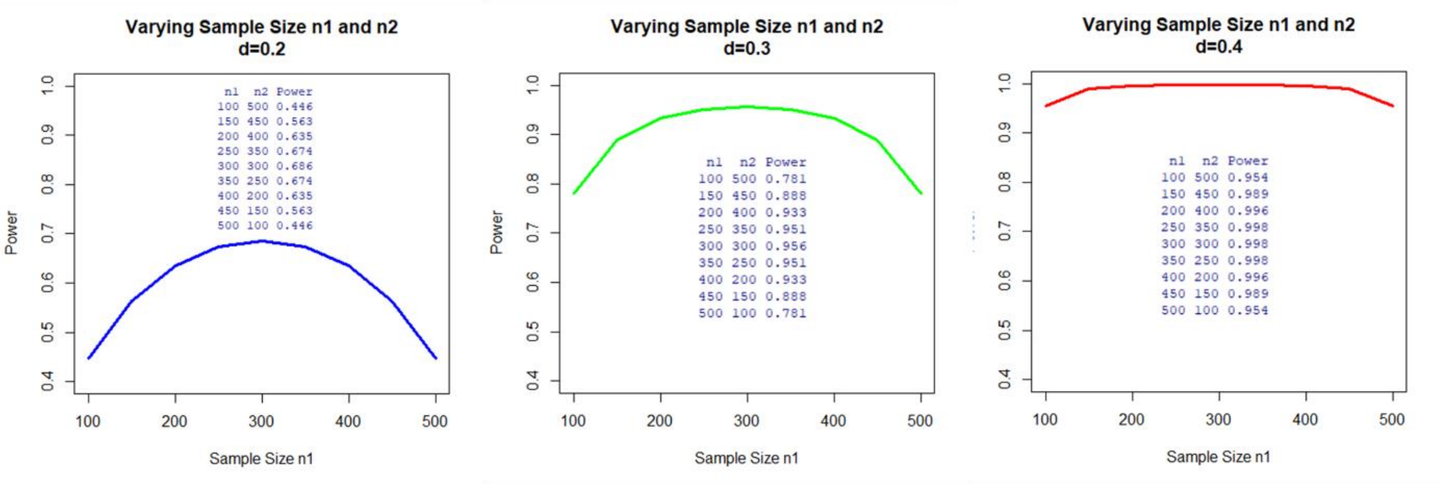
```
library(pwr)
pwr.t2n.test(d=0.2,n1=400,n2=250,
             alternative="two.sided")
```



```
 t test power calculation 

          n1 = 400
          n2 = 250
           d = 0.2
   sig.level = 0.05
       power = 0.6974449
 alternative = two.sided
```

sample size of 600, generating three scenarios to give three notings for a research study that intends to use a setting with unequal sample size. First, the peak in power is when the ratio is equal to one. Second, the symmetry pattern of the ratio is observed. Third, the higher the effect size, the less variation in power for unequal ratios. For the effect size set at .2, the power value is at a peak of .686 and drops to .446 when the ratio is either 1:5 or 5:1. When the effect size increases to 0.3, the differences in the magnitude in power between equal ratio and extreme ratio reduces. The impact on power becomes quite small for the effect size of 0.4. The systematic property of power to the sample ratio indicates when planning for a study, trying to avoid an unequal sample size ratio is always preferred to achieve a higher power especially when the expected effect size is small. When the effect size is large, the influence of unequal sample size on power becomes a trivial one.

Figure 4. Power by Varying Sample Size Ratio and Effect Size of Unbalanced Independent t-test



```
 n1  n2 Power
100 500 0.446
150 450 0.563
200 400 0.635
250 350 0.674
300 300 0.686
350 250 0.674
400 200 0.635
450 150 0.563
500 100 0.446
```

```
 n1  n2 Power
100 500 0.781
150 450 0.888
200 400 0.933
250 350 0.951
300 300 0.956
350 250 0.951
400 200 0.933
450 150 0.888
500 100 0.781
```

```
 n1  n2 Power
100 500 0.954
150 450 0.989
200 400 0.996
250 350 0.998
300 300 0.998
350 250 0.998
400 200 0.996
450 150 0.989
500 100 0.954
```

Welch's t-test – Unequal Variances

When equal variances in the population are violated, $\sigma_1^2 \neq \sigma_2^2$, Welch's t-test is generally suggested (Delacre, Lakens, & Leys, 2017; Fagerland & Sandvik, 2009; Welch, 1937). Power analysis that considered unequal standard deviation thus also developed. The function power.welch.t.test from the package MKPower provides the power analysis for Welch's t-test. The following shows three examples of specifying the standard deviations for sample 1 (sd1) and standard deviation for sample 2 (sd2) as (sd1, sd2). The ratio of the sd for the three examples specifies the two standard deviations as (1, 1), (0.5, 1.5), and (0.1, 2) respectively. The power values for the three specifications produce .87, .77, and .56 respectively.

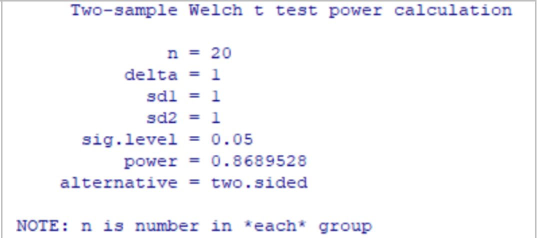
The three examples below demonstrate that the ratio of sd1 and sd2 do affect the level of power. The three power graphs in Figure 5 further illustrate the changes in the power by setting the delta with values 5, 6, and 7 and varying the ratio of the two standard deviations from 1:9 to 9:1. The first observation is that when the two standard deviations have the same value, the value of power is at the peak and reduces when the ratio of standard deviations is further apart. The symmetrical power by inverting the ratio is also observed. The third observation is the higher the value of delta, the effect on the power is lower. The practical implication is that the more deviate in the characteristic of the two samples, the lower is the impact on power even if the two standard deviations differ.

Bayesian Approach Using Bayes Factor

The Neyman–Pearson hypothesis testing has a long history of more than 50 years, however, numerous articles have criticized the frequentist approach in hypothesis testing (e.g., Wagenmakers, 2007, and Hubbard & Lindsay, 2008). While the pioneer Cohen (1992, 1988) play an important role in the earlier development of effect sizes and power analysis under the frequentist approach, and the development is mainly on the frequentist approach, the movement towards using the Bayesian approach for carrying out power analysis is growing (de Santis, 2007; Joseph et al., 2008; M'Lan et al., 2006; Rahme & Joseph, 1998; Shen et al, 2021; Shen et al, 2022; Simon, 1999; Wang & Gelfand, 2002) especially new free add-on R packages like SDDbain is made available with simple syntax specification to carry out Bayesian power analysis. This shift is also applied to the application of Bayesian power analysis for the t-test.

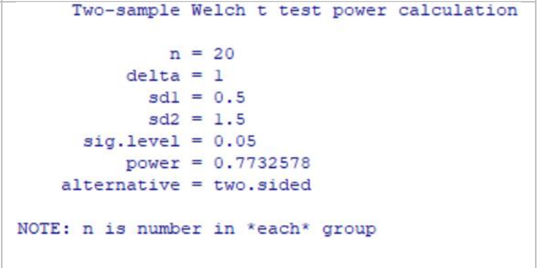
The central idea of Bayesian inferences is that a priori beliefs are updated with observed evidence and both are combined into a posterior distribution. The posterior distribution is used to predict subsequent

```
library(MKpower)
power.welch.t.test(n = 20, delta = 1,
  sd1=1.0, sd2=1.0)
```



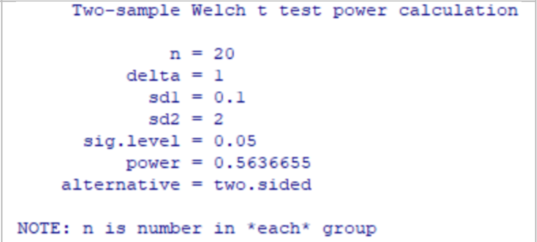
```
     Two-sample Welch t test power calculation 

              n = 20
          delta = 1
            sd1 = 1
            sd2 = 1
      sig.level = 0.05
          power = 0.8689528
    alternative = two.sided

NOTE: n is number in *each* group
```

```
library(MKpower)
power.welch.t.test(n = 20, delta = 1,
  sd1=0.5, sd2=1.5)
```



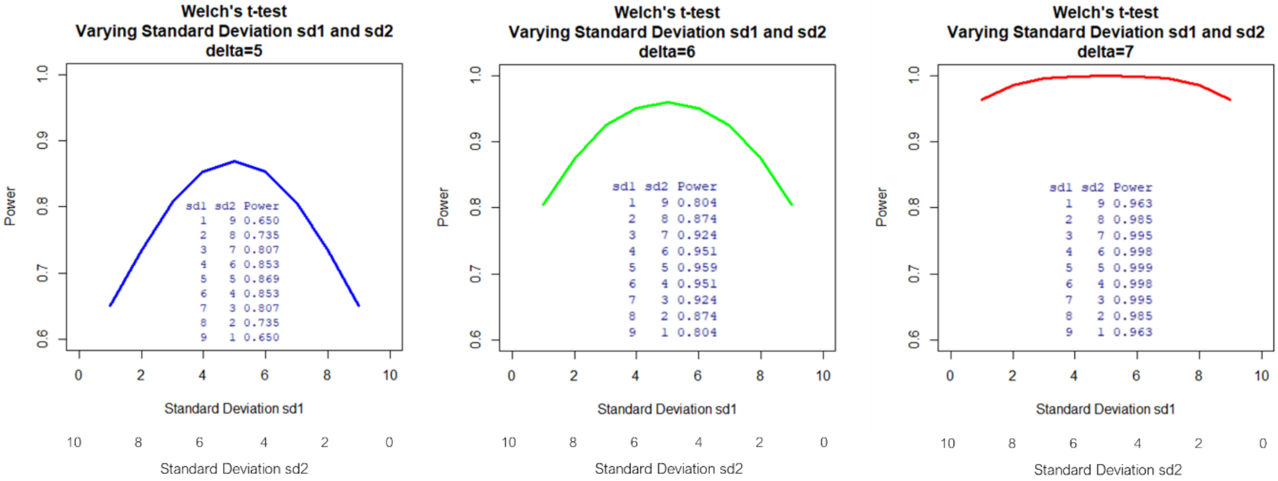
```
     Two-sample Welch t test power calculation 

              n = 20
          delta = 1
            sd1 = 0.5
            sd2 = 1.5
      sig.level = 0.05
          power = 0.7732578
    alternative = two.sided

NOTE: n is number in *each* group
```

```
library(MKpower)
power.welch.t.test(n = 20, delta = 1,
  sd1=0.1, sd2=2.0)
```



```
     Two-sample Welch t test power calculation 

              n = 20
          delta = 1
            sd1 = 0.1
            sd2 = 2
      sig.level = 0.05
          power = 0.5636655
    alternative = two.sided

NOTE: n is number in *each* group
```

Figure 5. Power by Varying Standard Deviation Ratio and Effect Size of Welch t-test



to show the procedure to generate sample size to evaluate t-test hypotheses using the approximate adjusted fractional Bayes factor (AAFBF) implemented in the R package bain (Gu et al, 2021) which carry out informative hypotheses using Bayes factors.

Bayes Factor

The Bayes factor (BF) is one of the oldest and most widely used indices for carrying out testing a hypothesis under the Bayesian framework. It compares the predictive ability of two competing models corresponding to both the hypotheses H_0 and H_1, indicating the degree of evidence a data set shifts the balance between the null hypothesis H_0 and the alternative hypothesis H_1 (Jeffreys, 1935). BF is a continuous measure of evidence for H_1 over H_0. When the Bayes factor is 1, the data is equally well predicted by both models, showing the evidence of not favoring either model over the other. When the BF increases above 1 the evidence favors H_1 over H_0. On the contrary, when the BF decreases below 1, the evidence favors H_0 over H_1 (Dienes and Mclatchie, 2018). More specifically, BF_{01} is the ratio of the two marginal likelihoods $p(data|H_0)$ and $p(data|H_1)$, each of which is calculated by integrating the respective model parameters according to their prior distribution (Kelter, 2021) as shown in Equation 1. It is noted $BF_{10} = 1/BF_{01}$ that specifies the evidence favors H_1 over H_0.

$$BF_{01} = \frac{p(data|H_0)}{p(data|H_1)}, BF_{10} = \frac{1}{BF_{01}} = \frac{p(data|H_1)}{p(data|H_0)} \quad (1)$$

Package SSDbain: Approximate Adjusted Fractional Bayes Factor and Informative Hypothesis

The advantage of using the Bayes factor becomes apparent over the incessant debate between frequentist and Bayesian hypothesis testing (Wagenmakers, 2007). However, it has suffered from two limitations. The specification of the prior can turn into a difficult task when prior information is weak or unavailable, and the computation can be very time-intensive. The package SSDbain (Fu, 2021) adopts the fractional Bayes factor (O'Hagan, 1995; O'Hagan, 1997) which is a partial Bayes factor method (de Santis & Spezzaferri, 1997, 1999) to address the limitation of using an appropriate prior distribution to determine the sample size using the approximate adjusted fractional Bayes factor (AAFBF) implemented in the R package bain (Gu et al, 2021). The basic idea of AAFBF is that a fraction of the data parameter is used to give the amount of information for specifying the prior as training, and the remaining fraction is used for testing the informative hypotheses. Another advantage of AAFBF is that instead of using an intrinsic Bayes factor (Berger & Pericchi, 1996) that is an average of the partial Bayes factors based on all possible minimal training samples that require more computer time, it partially reduces the time-intensive issue even though the time taken can still problematic when the number of data sets sampled from the null and alternative populations to determine the required sample is large.

Another feature of the package SSDbain is that an informative hypothesis is made possible. An informative hypothesis goes beyond the traditional hypothesis setting that could consist of equality and/or inequality constraints among the parameters of interest as well as the ordering of group means in order. The major advantage of evaluating a set of informative hypotheses using Bayesian model selection is that prior information can be incorporated into the analysis. Another advantage of evaluating informative hypotheses is that more power is generated with the same sample size (Fu et al, 2021).

Function SSDttest

This current paper introduces the function SSDttest R from the package SSDbain (Fu, Hoijtink & Moerbeek, 2021) to generate the sample size requirement for carrying out Bayesian power analysis for the Bayesian t-test and Welch's t-test. The function SSDttest generates the required sample size based on the updated priors following the fractional Bayes methodology as fractional priors. The specification of the t-test and Welch's t-test model are specified in Equations 2 and 3 respectively (Fu et al, 2021).

$$y_p = \mu_1 \mathrm{D}_{1p} + \mu_2 \mathrm{D}_{2p} + \epsilon_p \quad \sim N(0, \sigma^2) \qquad (2)$$

$$y_p = \mu_1 \mathrm{D}_{1p} + \mu_2 \mathrm{D}_{2p} + \epsilon_p \quad \sim N\left(0, \mathrm{D}_{1p}\sigma_1^2 + \mathrm{D}_{2p}\sigma_2^2\right) \qquad (3)$$

where $\mathrm{D}_{1p} = 1$ for person $p = 1, \dots, N$ and 0 otherwise, $\mathrm{D}_{2p} = 1$ for person $p = N+1, \dots, 2N$, and 0 otherwise, N denotes the common sample size for groups 1 and 2, ϵ_p denotes the error in prediction, σ^2 denotes the common within-group variance for both groups 1 and 2, and σ_1^2 and σ_2^2 denote the different within-group variances for groups 1 and 2, respectively. The prior distributions for the t-test and Welch's t-test are stated below for the former has a common variance in Equation 4 and the latter with different variances in Equation 5 (Gu et al, 2018; Hoijtink et al, 2019).

$$h_1(\boldsymbol{\mu}|\mathbf{y},\mathbf{D_1},\mathbf{D_2}) = N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\frac{1}{b}\frac{\hat{\sigma}^2}{N} & 0\\ 0 & \frac{1}{b}\frac{\hat{\sigma}^2}{N}\end{bmatrix}\right) \tag{4}$$

$$h_1(\boldsymbol{\mu}|\mathbf{y},\mathbf{D_1},\mathbf{D_2}) = N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\frac{1}{b}\frac{\hat{\sigma}_1^2}{N} & 0\\ 0 & \frac{1}{b}\frac{\hat{\sigma}_2^2}{N}\end{bmatrix}\right) \tag{5}$$

where b is the fraction of information in the data used to specify the prior distribution.

Six Scenarios – The Bayesian approach

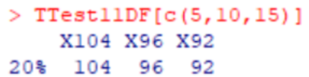
There are seven arguments under the function SSDttest that needs to specify to carry out sample size calculation. The first argument, type, is to specify whether it is a t-test or a Welch t-test. If type='equal', the Bayesian t-test is used; if type='unequal', the Bayesian Welch's t-test is specified. The population_mean argument specifies the population means of the two groups under H_1 and H_2. The third argument var indicates the two within-group variances. The eta argument specifies the probability that the Bayes factor is larger than the BFthresh if either the null hypothesis or the alternative hypothesis is true. The Hypothesis argument specifies the hypothesis. When Hypothesis='two-sided' is specified, the hypotheses are $H_0\colon \mu_1 = \mu_2$ and the alternative hypothesis $H_1\colon \mu_1 \neq \mu_2$. Hypothesis='one-sided' states the null hypothesis $H_0\colon \mu_1 = \mu_2$ and the alternative hypothesis $H_1\colon \mu_1 > \mu_2$. The last argument T specifies the number of datasets sampled from the null and alternative populations (Fu, 2021).

The following provides 6 scenarios to illustrate the use of package SSDbain, function SSDttest by varying the BF threshold, and η to illustrate their effects on sample size for both Bayesian t-test and Bayesian Welch's t-test. Scenarios 1 to 3 are catered for the Bayesian t-test and Scenarios 4 to 6 for the Bayesian Welch's t-test. Scenario 1 sets the base reference and generates the Bayesian t-test for carrying out a two-sided test. The specification of this base reference sets BFthresh to 3 and $\eta = 0.8$. The setting of Scenario 4 is the same as Scenario 1 but for Bayesian Welch's t-test. This is carried out by changing the argument type = "unequal" and specifying the two unequal variances by stating the argument Var = c(4/3,1/3). Scenarios 2 and 5 reduce the BFthresh to 2 and Scenario 3 and 6 increase η to 0.9 respectively for the Bayesian t-test and Welch's t-test.

The following provides the syntax for generating the sample size of the 6 scenarios. The full syntax is given for the first three scenarios. For the three corresponding Bayesian Welch's t-tests, scenarios 4 to 6, only the changes in syntax are stated. The results of Scenario 1 is the researcher should execute their study using the sample between 92 and 104. For Scenario 4, the number of sample sizes required is between 107 and 123.

Scenario 1

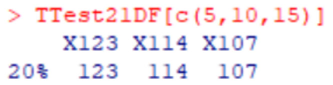
```
library(SSDbain)
set.seed(1234567)
TTest11 <- SSDttest(type='equal',
  Population_mean = c(0.5,0),
  var             = NULL,
  BFthresh        = 3,
  eta             = 0.8,
  Hypothesis      = 'two-sided',
  T               = 10000)
TTest11DF <- data.frame(TTest11)
TTest11DF[c(5,10,15)]
```



```
> TTest11DF[c(5,10,15)]
    X104 X96 X92
20%  104  96  92
```

Scenario 4

```
TTest21 <- SSDttest(type='unequal', ...
  Var = c(4/3,1/3),
...
```

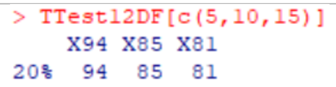
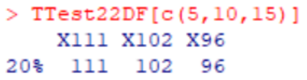


```
> TTest21DF[c(5,10,15)]
    X123 X114 X107
20%  123  114  107
```

By reducing the BFthresh from 3 to 2, the required sample size for Scenario 2 reduces to between 81 and 94 and for Scenario 5, the sample size required is between 96 to 111.

Scenario 2	Scenario 5

`library(SSDbain)`
`set.seed(1234567)`
`TTest12 <- SSDttest(type='equal',`
`  Population_mean = c(0.5,0),`
`  var             = NULL,`
`  BFthresh        = 2,`
`  eta             = 0.8,`
`  Hypothesis      = 'two-sided',`
`  T               = 10000)`
`TTest12DF <- data.frame(TTest12)`
`TTest12DF[c(5,10,15)]`
	`TTest22 <- SSDttest(type='unequal', ...`
	`  Var = c(4/3,1/3),`
	`...`

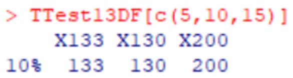
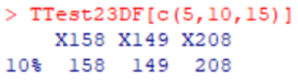

`> TTest12DF[c(5,10,15)]`
`    X94 X85 X81`
`20%  94  85  81`
	
	`> TTest22DF[c(5,10,15)]`
	`    X111 X102 X96`
	`20%  111  102  96`

With the η increases from 0.8 to 0.9, the required sample increases to between 130 to 200 for Scenario 3 and between 149 to 208 for Scenario 6.

In summary, when the BF threshold decreases, the required sample decreases. On the contrary, an increase in η will accompany an increase in sample size. The required sample size for Welch's t-test is higher than the t-test since the variances are unequal.

Scenario 3	Scenario 6

`library(SSDbain)`
`set.seed(1234567)`
`TTest13 <- SSDttest(type='equal',`
`  Population_mean = c(0.5,0),`
`  var             = NULL,`
`  BFthresh        = 3,`
`  eta             = 0.9,`
`  Hypothesis      = 'two-sided',`
`  T               = 10000)`
`TTest13DF <- data.frame(TTest13)`
`TTest13DF[c(5,10,15)]`
	`TTest23 <- SSDttest(type='unequal', ...`
	`  Var = c(4/3,1/3),`
	`...`


`> TTest13DF[c(5,10,15)]`
`    X133 X130 X200`
`10%  133  130  200`
	
	`> TTest23DF[c(5,10,15)]`
	`    X158 X149 X208`
	`10%  158  149  208`

Summary and Conclusions

The procedure of choosing an appropriate sample size using the various R packages is the main focus of this paper. Both the frequentist and the Bayesian approaches to carrying out power analysis are illustrated with syntax, and examples and the implications of using the various approaches and practical concerns on the determination of sample size, power, and effect size are discussed. Table 1 summarises the functions to generate power analysis from the R Base functions (R Core Team, 2021), package pwr (Champely, 2020), package MKPower (Kohl, 2020), package WebPower (Zhang & Mai, 2021; Zhang & Yuan, 2018), package MKpower (Kohl, 2020), and package SSDbain (Fu, 2021).

For one sample t-test, this paper provides a written function and an R program to examine the sensitivity by varying a factor to examine the output of power analysis. The function PowerCurve.OneSample.FixedN plots the effect of the power given a numeric input of sample size, and the R program generates a power curve by varying the sample and effect size given the level of power. Both power curves show that the most sensitive factor affecting the sample size is the effect size. The same results could also be determined for the paired t-test power analysis. The practical concern is that if a researcher is unaware of the size of the effect, effect size becomes the crucial factor that has to be taken more seriously for the determination of sample size. Examining the power curve provides a useful way for choosing an appropriate sample size and examining the sensitivity of factors affecting the sample size. In practice, it is often found sample sizes could differ for two independent samples. A researcher that

takes note of the effect of sample size for unbalanced design on the power will help to decide on the selection of an appropriate sample. A similar phenomenon could also be observed for Welch's t-test when the ratio of two variances differs. The implication of unequal variances to power analysis about effect size also could help a researcher to make a reasonable decision when faced with unequal variances.

The six scenarios of the two-sample Bayesian power analysis show the evidence that the higher the deviation from equal variances, the higher the specified η, the higher the BF setting, and the higher the required sample size.

Table 1. Functions for Power Analysis

Function	**Description**
Base	
power.t.test(type="one.sample")	Power calculation for one sample t-test
power.t.test(type="one.sample")	Power calculation for independent t-test
power.t.test(type="paired")	Power calculation for paired t-test
Package pwr	
pwr.t.test(type="one.sample")	Power calculation for one sample t-test
pwr.t.test(type="two.sample")	Power calculation for independent t-test
pwr.t.test(type="paired")	Power calculation for paired t-test
pwr.t2n.test()	Power calculations for two samples with unequal n t-test
Package WebPower	
wp.t(type="one.sample")	Power calculation for one sample t-test
wp.t(type="two.sample")	Power calculation for independent t-test
wp.t(type="paired")	Power calculation for paired t-test
wp.t(type="two.sample.2n")	Power calculations for two samples with unequal n t-test
Package MKpower	
power.welch.t.test(sd1=,sd2=)	Power calculation for Welch's t-test of unequal variance
Package SSDbain	
SSDttest(type="equal")	Power calculation for Bayesian t-test using Bayes Factor
SSDttest(type="unequal")	Power calculation forBayesian Welch's t-test using Bayes Factor

Discussion

The importance of carrying out power analysis is clearly stated by Cohen (1988) that estimating the expected statistical power before beginning research by power analysis is crucial to avoid making a wrong conclusion. The mathematical equations provided by Cohen (1988, 1992) give the relation between effect size, sample size, type I error rate, and power. Setting the power at 80%, with type I error rate α specifies as .05, the minimum sample size per group are 394, 64, and 26 for small (d = 0.2), medium (d = 0.5), and large (d = 0.8) effect sizes respectively for an independent sample two-sided t-test. This rule of thumb becomes almost a standard reference for carrying out frequentist power analysis. However, in practice, sample size determination is not a straightforward statistical computation issue. The rule of thumb set by the authority such as Cohen's recommendation for the broad effect sizes categories of the small, medium, and large to cover the range of effect sizes are commonly referred to as a general guide. However, other practical factors could go beyond statistical issues that affect the determination of sample size. Dong & Maynard (2013) distinguish two types of factors: discretionary and inherent factors. Discretionary factors are factors based on the researcher's judgment and statistical-based decision. Inherent factors, on the other hand, are factors outside of the researcher's control that depend on the nature of the intervention, the study design, and the characteristics of the true effect.

Most often, the best first response to answer an appropriate sample size for a t-test is not a number, but a sequence of relevant questions to ascertain the ultimate research concern on the expected effect. A study's size and structure depend on the research context, including the researcher's objectives and proposed analyses. This goes back to the fundamental concern of carrying out power analysis is that a well-designed research project has to take account of the study to have a meaningful size impact, that there is a high probability the study will detect it and not too big a sample size in wasting unnecessarily sources to collect extra data (Legg & Nagy 2006). For instance, a research study that aims to use a more sophisticated analytic technique may require a larger sample than simple techniques. When unequal variances of two samples need to be considered, the required analytical requirements not only differ from the usual t-test but the unequal variances need to know beforehand to determine the impact to derive the sample size. Most often, in practice, due to cost constraints, achieving a level of power is almost practically impossible. Determining the sample size thus goes beyond using a derived formula but to carrying out the necessary sensitivity analysis to determine a whole range of sample sizes that take into consideration the inherent factors that could affect the sample size.

For Bayesian power analysis, the nature and the importance of the research could be considered in determining the sample size, the flexibility lies in the specification of the value of BF. A large BF thresh value for high-stakes research and a lower BF for a lower stake study is one of the guidelines one could adopt. Similarly, a high value of η such as .90 for high-stake research and reduce accordingly for lower-stake study. For instance, in pharmaceutical research, a new headache relief drug compared to an existing competitor could specify a low BF whereas, for cancer and COVID research, a high BF is generally expected (Fu et al, 2021).

Choosing between using the frequentist and Bayesian approach is still an open debate. From the Bayesian perspective, unknown parameters are random variables following certain distributions. Therefore, developing a Bayesian method assuming the parameters are random variables to resolve the issue of uncertainty. But, the choosing of a BF threshold and η can be equally subjective. Researchers may have to refer to the general recommendation of common values for BF to set at 3, 5, or 10. Probably, the greatest disadvantage of using the Bayesian approach is that the time taken can be far too long to generate the output. Fu et al (2021) recommend the number of datasets to perform the SSDttest be to set at least 10,000. This can take some time, using a PC notebook.

References

Berger, J. O. & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association, 91(433),* 109-122.

Carneiro, A. V. (2003). Estimating sample size in clinical studies: basic methodological principles.

Revista Portuguesa de Cardiologia, 22(12),1513-1521. English, Portuguese. PMID: 15008067.

Champely, S. (2020). *pwr: Basic functions for power analysis. R package version 1.3-0.* https://CRAN.R-project.org/package=pwr

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). New York: Routledge.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112(1)*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155.

Connelly, L. M. (2008). Research considerations: Power analysis and effect size. *MedSurg Nursing, 17(1)*, 41-42.

Datanami (2020). *Left for Dead, R Surges Again.* https://www.datanami.com/2020/07/10/left-for-dead-r-surges-again/.

de Santis, F., & Spezzaferri, F. (1997). Alternative Bayes factors for model selection. *Canadian Journal of Statistics, 25*, 503–515. https://doi.org/10.2307/3315344.

de Santis, F., & Spezzaferri, F. (1999). Methods for default and robust Bayesian model comparison: The fractional Bayes factor approach. *International Statistical Review, 67*, 1–20. https://doi.org/10.2307/1403706.

de Santis, F. (2007). Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170(1),* 95–113.

Delacre, M, Lakens, D. & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology, 30(1),* 92-101, DOI: https://doi.org/10.5334/irsp.82.

Desu, M. M. & Raghavarao, D. (1990). *Sample Size Methodology*. Boston, Academic Press.

Dienes, Z. & Mclatchie, N. (2018). Four reasons to prefer Bayesian analysis over significance testing. *Psychonomic Bulletin & Review, 25*, 207–218. DOI: 10.3758/s13423-017-1266-z.

Dong, N & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi experimental design. Journal of Research on Educational Effectiveness, 6(1), 24-67. DOI: 10.1080/19345747.2012.673143.

Fagerland, M W. & Sandvik, L. (2009). The Wilcoxon-Mann-Whitney test under scrutiny. *Statistics in Medicine, 28*, 1487-1497.

Fu, Q. (2021). *SSDbain: Sample size determination using AAFBF for Bayesian hypothesis testing implemented in bain. R package version 0.1.0.*

Fu, Q. Hoijtink, H. & Moerbeek, M. (2021). Sample-size determination for the Bayesian t-test and Welch's test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods, 53*, 139–152. https://doi.org/10.3758/s13428-020-01408-1.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78.

Gu, X., Mulder, J. & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology, 71*, 229–261. DOI: 10.1111/bmsp.12110.

Gu, X., Hoijtink, H., Mulder, J. & van Lissa. C. J. (2021). *bain: Bayes Factors for Informative Hypotheses. R package version 0.2.8.* https://CRAN.R-project.org/package=bain.

Hoijtink, H., Gu, X., & Mulder, J. (2019). Bayesian evaluation of informative hypotheses for multiple populations. *British Journal of Mathematical and Statistical Psychology, 72(2)*, 219–243. https://doi.org/10.1111/bmsp.12145.

Hubbard, R., & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18(1)*, 69–88. https://doi.org/10.1177/0959354307086923.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society, 31(2)*, 203–222. https://doi.org/10.1017/S030500410001330X.

Kohl, M. (2020). *MKpower: Power analysis and sample size calculation. R package version 0.5*, URL: http://www.stamats.de.

Legg, C. J. & Nagy, L. (2006). Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management, 78*, 194-199.

Livingston, E. H. & Cassidy, L. (2005) Statistical power and estimation of the number of required subjects for a study based on the t-test: A surgeon's primer. *Journal of Surgical Research, 126*, 149–159. DOI:10.1016/j.jss.2004.12.013.

Lovakov, A., & Agadullina, E. R. (2021). Empirically Derived Guidelines for Effect Size Interpretation in Social Psychology. *European Journal of Social Psychology, 51*, 485-504. DOI: 10.1002/ejsp.2752.

M'Lan C. E. & Joseph, L. & Wolfson, D. B. (2006). Bayesian sample size determination for case-control studies. *Journal of the American Statistical Association, 101(474)*, 760–772.

Nuijten, N. B., Hartgerink, C. H. J., van Assen, M. A. L. M, & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods, 48(4)*, 1205-1226. DOI 10.3758/s13428-015-0664-2.

Joseph, L., M'Lan, C. E. & Wolfson, D. B. (2008). Bayesian sample size determination for binomial proportions. *Bayesian Analysis, 3*, 2.

O'Hagan, A. (1995). Fractional Bayes factors for model comparisons (with discussion). *Journal of the Royal Statistical Society, Series B, 57*, 99–138.

O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factors. *Test, 6,* 101–118. https://doi.org/10.1007/BF02564428.

R Core Team (2021*). R: A language and environment for statistical computing.* R *Foundation for Statistical Computing*, Vienna, Austria. URL: https://www.R-project.org/.

Stack Overflow Blog. (2017). *The Impressive Growth of* R. https://stackoverflow.blog/2017/10/10/impressive-growth-r/.

Rahme, E., Joseph, L. (1998). Exact sample size determination for binomial experiments. *Journal of Statistical Planning and Inference, 66*, 83–93.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with* R. Springer, New York.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods, 8(2)*, 597-599.

Shen, Y., Psioda, M. A., & Ibrahim, J. G. (2022). *BayesPPD: Bayesian Power Prior Design.* R *package version 1.0.4.* https://CRAN.R-project.org/package=BayesPPD.

Shen, Y., Psioda, M. A., & Ibrahim, J. G. (2021). *BayesPPD: An* R *package for Bayesian sample size determination using the power and normalized power prior for generalized linear models.* Department of Biostatistics, University of North Carolina at Chapel.

Simon R. (1999). Bayesian design and analysis of active control clinical trials. *Biometrics, 55(2)*, 484–487.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14(5),* 779–804. https://doi.org/10.3758/BF03194105.

Wang, F. & Gelfand, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science, 17(2),* 193–208.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29*, 350–362.

Wickham. H.(2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Zhang, Z, & Yuan, K.-H. (2018). *Practical Statistical Power Analysis using WebPower and* R. ISDSA Press, Granfer, IN.

Zhang, Z. & Mai, Y. (2021). *WebPower: Basic and advanced statistical power analysis.* R *package version 0.6.* https://CRAN.R-project.org/package=WebPower.

Citation:
Tan, T. K.. (2022). Practical t-test power analysis with R. *Practical Assessment, Research, & Evaluation*, 27(18). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/18/

Corresponding Author:
Teck Kiang Tan
National University of Singapore
Singapore

Email: alsttk [at] nus.edu.sg

Appendix A.
Graphing Power by Fixing Sample Size and Graphing Sample Size Required by Fixing Power for One-Sample t-test

A-1.Function PowerCurve.OneSample.FixedN

```
library(pwr)
library(ggplot2)
library(ggrepel)
PowerCurve.OneSample.FixedN <- function(n){
  ES <- seq(.1,0.9,.1)  # Vector of effect size
  samp.out <- NULL
  for(i in 1:length(ES)){
    power <-  pwr.t.test(d=ES[i],n=n,sig.level=.05,type="one.sample")$power
    power <-  data.frame(effect.size=ES[i],power=power)
    samp.out <- rbind(samp.out,power)
  }
  print(samp.out)
  ggplot(samp.out, aes(effect.size,power,label=effect.size))+
    geom_line(color="pink") +
    geom_point(color="red") +
    geom_text_repel(color="red") +
    theme(plot.title=element_text(hjust=0.5,size=15),
          plot.subtitle=element_text(hjust=0.5,size=12)) +
    geom_hline(yintercept = .8,lty=2, color='blue', size=1.1) +
    labs(title = paste0("t-test Power Curve for n=", n),
         subtitle = "Effect Size from 0.1 to 0.9 by 0.1",
         x     = "Cohen's d",
         y     = "Power")
}
```

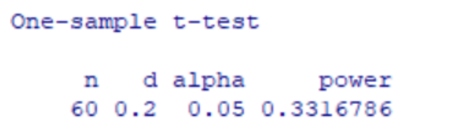
A-2. Function PowerCurve.OneSample.FixedPower

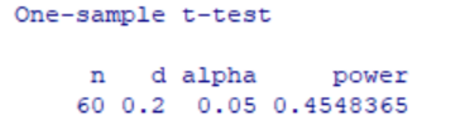
```
library(pwr)
library(ggplot2)
library(ggrepel)
PowerCurve.OneSample.FixedPower <- function(p){
  ES <- seq(.2,0.9,.1)  # Vector of effect size
  samp.out <- NULL
  for(i in 1:length(ES)){
    sample.size <-  pwr.t.test(d=ES[i],power=p,sig.level=.05,type="one.sample")$n
    sample.size <-  data.frame(effect.size=ES[i],sample.size)
    samp.out <- rbind(samp.out,sample.size)
  }
  print(samp.out)
  ggplot(samp.out, aes(sample.size, effect.size, label=round(sample.size)))+
    geom_line(color="pink") +
    geom_point(color="red") +
    geom_text_repel(color="red") +
    theme(plot.title=element_text(hjust=0.5,size=15),
          plot.subtitle=element_text(hjust=0.5,size=12)) +
    labs(title = paste0("One Sample t-test\nVary Sample & Effect Size for Power = ", p),
         subtitle = "Effect Size from 0.2 to 0.9 by 0.1",
         y     = "Effect Size",
         x     = "Sample Size")
}
p <- 0.8
PowerCurve.OneSample.FixedPower(p)
p <- 0.6
PowerCurve.OneSample.FixedPower(p)
library(ggplot2)
library(ggrepel)
PowerCurve.OneSample.FixedPower <- function(p){
  ES <- seq(.2,0.9,.1)  # Vector of effect size
  samp.out <- NULL
  for(i in 1:length(ES)){
    sample.size <-  pwr.t.test(d=ES[i],power=p,sig.level=.05,type="one.sample")$n
    sample.size <-  data.frame(effect.size=ES[i],sample.size)
    samp.out <- rbind(samp.out,sample.size)
  }
  print(samp.out)
  ggplot(samp.out, aes(sample.size, effect.size, label=effect.size))+
    geom_line(color="pink") +
    geom_point(color="red") +
    geom_text_repel(color="red") +
    theme(plot.title=element_text(hjust=0.5,size=15),
          plot.subtitle=element_text(hjust=0.5,size=12)) +
    geom_hline(yintercept = .8,lty=2, color='blue', size=1.1) +
    labs(title = paste0("t-test Power Curve for Power = ", p),
         subtitle = "Effect Size from 0.2 to 0.9 by 0.1",
         y     = "Power",
         x     = "Sample Size")
}
p <- 0.8
PowerCurve.OneSample.FixedPower(p)
p <- 0.6
PowerCurve.OneSample.FixedPower(p)
```

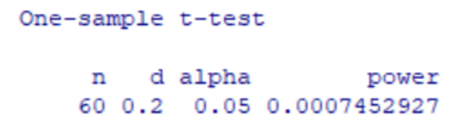
Appendix B.
Package WebPower, Function wp.t

The R syntax to duplicate the power analysis in the main text for the one-sample, two-sample paired t-test, two-sample independent, and unbalanced independent t-test are printed below using the function wp.t under the package WebPower package. The plot function generates the power curve for the one-sample t-test also provided.

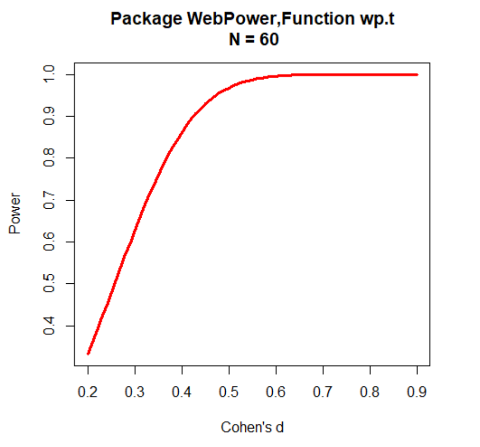
Table B-1. One-Sample t-test – Estimate Power

`library(WebPower)`
`wp.t(n1=60, d=0.2, type="one.sample")`
	`One-sample t-test`
	`  n   d alpha     power`
	`60 0.2  0.05 0.3316786`

`wp.t(n1=60, d=0.2, type="one.sample",`
`     alternative = c("greater"))`
	`One-sample t-test`
	`  n   d alpha     power`
	`60 0.2  0.05 0.4548365`

`wp.t(n1=60, d=0.2, type="one.sample",`
`     alternative = c("less"))`
	`One-sample t-test`
	`  n   d alpha        power`
	`60 0.2  0.05 0.0007452927`

```
Plot1 <- wp.t(n1=60, d=seq(0.2,0.9,0.01),
              type="one.sample")
plot(Plot1$d,Plot1$power,
 type = "l",col = "red",lwd = 3,
 main = "Package WebPower,Function wp.t\n N = 60",
 xlab = "Cohen's d",
 ylab = "Power")
```



```
Plot2 <- wp.t(n1=100, d=seq(0.2,0.9,0.01),
              type="one.sample")
plot(Plot2$d,Plot2$power,
 type = "l",col = "blue",lwd = 3,
 main = "Package WebPower,Function wp.t\n N = 100",
 xlab = "Cohen's d",
 ylab = "Power")
```

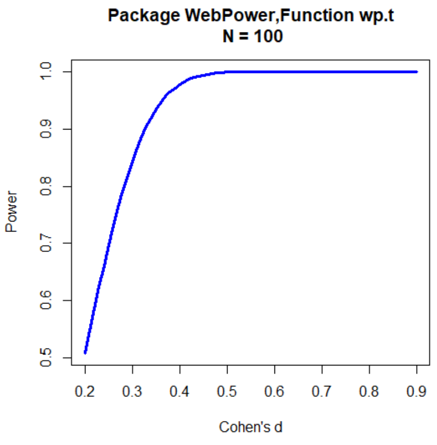
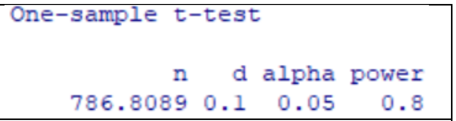


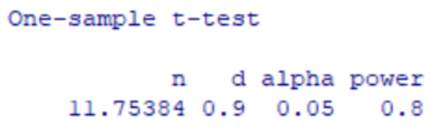
Table B-2. One-Sample t-test – Estimate Sample Size

```
library(WebPower)
wp.t(d=0.1, type="one.sample",
     power=0.8,
     alternative = c("two.sided"))
```



```
One-sample t-test

            n   d alpha power
     786.8089 0.1  0.05   0.8
```

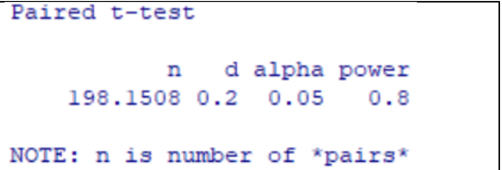
```
wp.t(d=0.9, type="one.sample",
     power=0.8,
     alternative = c("two.sided"))
```



```
One-sample t-test

            n   d alpha power
     11.75384 0.9  0.05   0.8
```

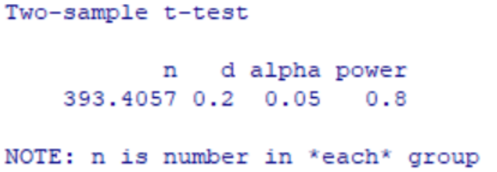
Table B-3. Two-Sample Paired and Independent t-test – Estimate Sample Size

```
library(WebPower)
wp.t(power=.8, d=0.2,
     type="paired",
     alternative="two.sided")
```



```
Paired t-test

            n   d alpha power
     198.1508 0.2  0.05   0.8

NOTE: n is number of *pairs*
```

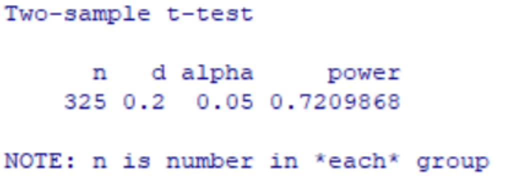
```
wp.t(power=.8, d=0.2,
     type="two.sample",
     alternative="two.sided")
```



```
Two-sample t-test

            n   d alpha power
     393.4057 0.2  0.05   0.8

NOTE: n is number in *each* group
```

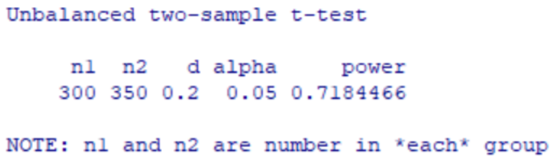
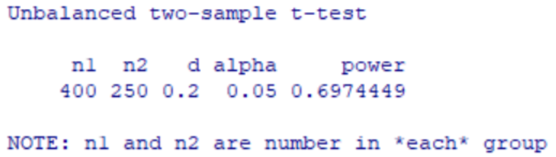
Table B-4. Two-Sample Independent t-test – Estimate Power

```
library(WebPower)
wp.t(d=0.2, n1=325,
     type="two.sample",
     alternative="two.sided")
```



```
Two-sample t-test

       n   d alpha     power
     325 0.2  0.05 0.7209868

NOTE: n is number in *each* group
```

Table B-5. Unbalanced Two-Sample Independent t-test – Estimate Power

library(WebPower) **wp.t(n1=300,n2=350,d=0.2,** **type="two.sample.2n",** **alternative="two.sided")**	 Unbalanced two-sample t-test n1 n2 d alpha power 300 350 0.2 0.05 0.7184466 NOTE: n1 and n2 are number in *each* group
wp.t(n1=400,n2=250,d=0.2, **type="two.sample.2n",** **alternative="two.sided")**	 Unbalanced two-sample t-test n1 n2 d alpha power 400 250 0.2 0.05 0.6974449 NOTE: n1 and n2 are number in *each* group

Appendix C.
R Program – Obtain Power Curve by Varying Effect Size and Power Value

```
library(pwr)
# Range of Effect Sizes
d <- seq(0.1,0.6,0.05)
nd <- length(d)

# Power Values
p <- seq(.6,.8,.05)
np <- length(p)

# Obtain Sample Sizes
samsize <- array(numeric(nd*np), dim=c(nd,np))
for (i in 1:np){
  for (j in 1:nd){
    result <- pwr.t.test(n = NULL, d = d[j],
    sig.level = .05, power = p[i], type = "paired",
    alternative = "two.sided")
    samsize[j,i] <- ceiling(result$n)
  }
}

# Plot Power Curve
xrange <- range(d)
yrange <- round(range(samsize))
colors <- rainbow(length(p))
plot(xrange, yrange, type="n",
  xlab="Effect Size (d)",
  ylab="Sample Size (n)" )

# add power curves
for (i in 1:np){
  lines(d, samsize[,i], type="l", lwd=2, col=colors[i])
}

# add annotation (grid lines, title, legend)
abline(v=0, h=seq(0,yrange[2],50), lty=2, col="grey89")
abline(h=0, v=seq(xrange[1],xrange[2],.02), lty=2,
   col="grey89")
title("Sample Size Estimation for Paired t-test\n
  Sig=0.05 (Two-tailed)")
legend("topright", title="Power", as.character(p),
   fill=colors)

# ---------------------- #
# Print Sample Size Table #
# ---------------------- #
ttestPairedN <- samsize
colnames(ttestPairedN) <- c(round(p,2))
rownames(ttestPairedN) <- c(round(d,2))
ttestPairedN
```

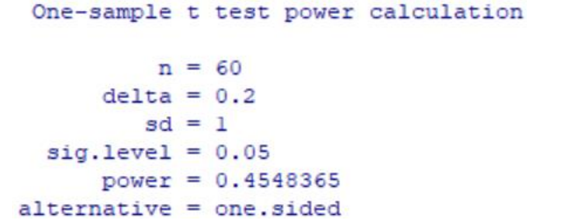
Effect Size	Power				
	0.60	0.65	0.70	0.75	0.80
0.10	492	552	620	696	787
0.15	220	247	277	311	351
0.20	125	140	157	176	199
0.25	81	90	101	113	128
0.30	57	64	71	80	90
0.35	42	47	53	59	67
0.40	33	37	41	46	52
0.45	27	30	33	37	41
0.50	22	24	27	30	34
0.55	19	21	23	25	28
0.60	16	18	20	22	24

Appendix D.
Base Function power.t.test

The R syntaxes and outputs using the base function power.t.test are printed below.

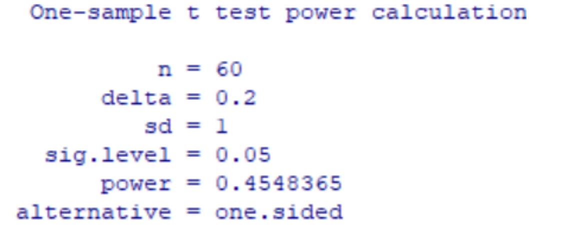
Table D-1. One-Sample t-test – Estimate Power

```
power.t.test(d           = 0.2,
             n           = 60,
             sig.level   = 0.05,
             type        = "one.sample",
             alternative = "two.sided")
```



```
  One-sample t test power calculation

            n = 60
        delta = 0.2
           sd = 1
    sig.level = 0.05
        power = 0.4548365
  alternative = one.sided
```

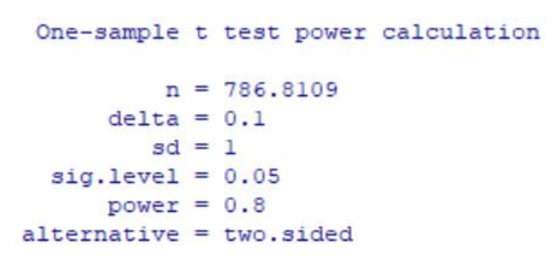
```
power.t.test(d           = 0.2,
             n           = 60,
             sig.level   = 0.05,
             type        = "one.sample",
             alternative = "one.sided")
```



```
  One-sample t test power calculation

            n = 60
        delta = 0.2
           sd = 1
    sig.level = 0.05
        power = 0.4548365
  alternative = one.sided
```

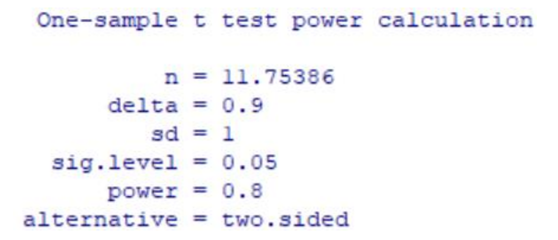
Table D-2. One-Sample t-test – Estimate Sample Size

```
power.t.test(d           = 0.1,
             power       = 0.8,
             sig.level   = 0.05,
             type        = "one.sample",
             alternative = "two.sided")
```



```
  One-sample t test power calculation

            n = 786.8109
        delta = 0.1
           sd = 1
    sig.level = 0.05
        power = 0.8
  alternative = two.sided
```

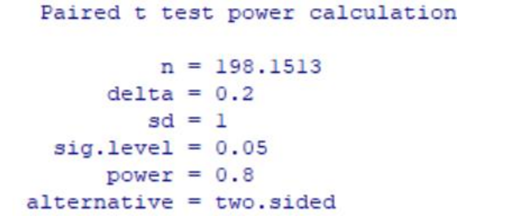
```
power.t.test(d           = 0.9,
             power       = 0.8,
             sig.level   = 0.05,
             type        = "one.sample",
             alternative = "two.sided")
```



```
  One-sample t test power calculation

            n = 11.75386
        delta = 0.9
           sd = 1
    sig.level = 0.05
        power = 0.8
  alternative = two.sided
```

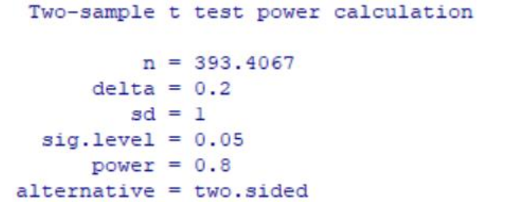
Table D-3. Two-Sample Paired and Independent t-test – Estimate Sample Size

```
power.t.test(d           = 0.2,
             power       = 0.8,
             sig.level   = 0.05,
             type        = "paired",
             alternative = "two.sided")
```



```
     Paired t test power calculation

              n = 198.1513
          delta = 0.2
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

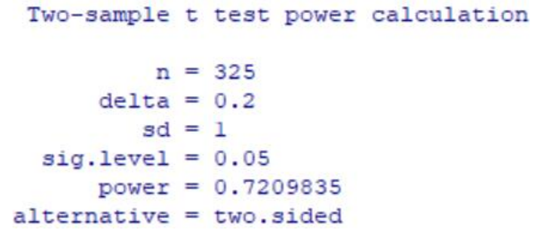
```
power.t.test(d           = 0.2,
             power       = 0.8,
             sig.level   = 0.05,
             type        = "two.sample",
             alternative = "two.sided")
```



```
     Two-sample t test power calculation

              n = 393.4067
          delta = 0.2
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Table D-4. Two-Sample Independent t-test – Estimate Power

```
power.t.test(d           = 0.2,
             n           = 325,
             sig.level   = 0.05,
             type        = "two.sample",
             alternative = "two.sided")
```



```
     Two-sample t test power calculation

              n = 325
          delta = 0.2
             sd = 1
      sig.level = 0.05
          power = 0.7209835
    alternative = two.sided
```

Appendix E.
Rule of Thumb – Effect Size Description of Cohen's d

Table E-1. Rule of Thumb – Effect Size Description of Cohen's d

Description	**Cohen (1988)**	**Sawilowsky (2009)**	**Lovakov & Agadullina (2021)**	**Gignac & Szodorai (2016)**
Tiny		d < 0.1		
Very Small	d < 0.2	0.1 <= d < 0.2	d < 0.15	d < 0.20
Small	0.2 <= d < 0.5	0.2 <= d < 0.5	0.15 <= d < 0.36	0.20 <= d < 0.41
Medium / Moderate	0.5 <= d < 0.8	0.5 <= d < 0.8	0.36 <= d < 0.65	0.41 <= d < 0.63
Large	d >= 0.8	0.8 <= d < 1.2	d >= 0.65	d >= 0.63
Very Lare		1.2 <= d < 2.0		
Huge		d >= 2		