


Probabilistic Thinking Is the Name of the Game: Integrating Test and Confidence Intervals to Plan Sample Sizes

Wei-Ming Luh¹ 

[1] *Institute of Education, National Cheng Kung University, Tainan, Taiwan.*

Methodology, 2022, Vol. 18(2), 80–98, <https://doi.org/10.5964/meth.6863>

Received: 2021-06-05 • **Accepted:** 2022-04-20 • **Published (VoR):** 2022-06-30

Handling Editor: Jost Reinecke, Bielefeld University, Bielefeld, Germany

Corresponding Author: Wei-Ming Luh, Institute of Education, National Cheng Kung University, #1 University Road, Tainan, 701, Taiwan. E-mail: luhwei@mail.ncku.edu.tw

Abstract

Having high statistical power and good estimated precision are essential to statistical practice; however, this integrative consideration on sample size planning remains limited in the literature, especially for two-group mean comparisons with unequal/unknown variances and unequal sampling costs. Furthermore, due to the neglect or misuse of employing confidence intervals, the present study aims to illuminate the probabilistic thinking by finding optimal allocations of sample sizes such that researchers can claim that the null hypothesis is rejected, the desired confidence-interval width of mean difference is achieved, and/or the true difference is encompassed in the interval. Cost effectiveness was also considered to find the optimal sample size. The simulation showed that the proposed approach can maintain the desired probability level for the conditional/unconditional probabilities of events and has good coverage rates in terms of confidence intervals. This study provides an important opportunity to advance the understanding of sample size planning and confidence intervals as well. Three R Shiny apps are provided for easy application in the Supplementary Materials.

Keywords

desired width, sampling cost, unconditional probability, unequal/unknown variances, Welch's t test

Sample size planning is a classic problem in research design when aiming for the greatest statistical power to reject the null hypothesis, while also having the smallest sample size in order to be economical. Moreover, to obtain confidence intervals (CIs) in data analysis is an increasingly important topic because CIs provide substantive advantag-



es to the controversial null hypothesis significance testing by providing informative results and facilitating the accumulation of knowledge from insufficient data (American Psychological Association, 2001; Cumming & Finch, 2001; Meeker et al., 2017; Mendoza & Stafford, 2001; Smithson, 2000; Thompson, 1998; also see the 2019 special issue of the *American Statistician*). Unfortunately, misunderstandings of CIs are not uncommon, and further guidance regarding their use and interpretation is still needed (Belia et al., 2005; Cumming & Calin-Jageman, 2017; Fidler et al., 2004; Morey et al., 2016). Namely, asking how much and how uncertain are the salient statistical literacy (Calin-Jageman & Cumming, 2019; Trafimow, 2018; Trafimow & Uhalt, 2020). In particular, it should be noted that the sample size calculation for constructing CIs may be different from those for testing hypotheses (Borenstein et al., 2001) because the expected width or precision is the concern, rather than the statistical power (Kelley et al., 2003). More specifically, a naïve researcher usually ignores the fact that CI width is actually a random variable; as such, the sample size might be underestimated, and the resulting CI would likely only have about 50% of probability of containing the true parameter (Grieve, 1989; Liu, 2009). In view of the problem mentioned, an integrative method for sample size calculation which can incorporate statistical power and precision is a continuing concern.

For comparing two-group means, recent developments in sample size planning have already heightened the need to integrate the notion of power and precision altogether by probabilistic thinking. An increasingly recognized solution to the integration is to reject the null hypothesis (event rejection, R), to encompass the true mean difference for a $100(1 - \alpha)\%$ two-sided CI (event validity, V), and to achieve the desired width for this CI (event width, W), (Beal, 1989; Cesana & Antonelli, 2010; Jiroutek et al., 2003; Kelley et al., 2003; Liu, 2009, 2012; Pan & Kupper, 1999). For the event rejection, the probability $P(R)$ is defined as the probability of the null hypothesis being rejected. For the event width, the probability $P(W)$ is the probability of the CI width being smaller than, or equal to, a desired width value. Note that this unconditional probability does not depend on whether the CI includes the true parameter value. Kelley and Rausch (2006) proposed a method for sample size planning via the accuracy in the parameter estimation (AIPE) approach for the expectation value of the width; thereafter, Lai and Kelley (2012) extended AIPE to mean contrasts. Moreover, Beal (1989) proposed a conditional probability, denoted by $P(W|V)$, which is the probability of a random CI width achieving a desired width value, conditional on the CI including the true parameter value. Furthermore, Jiroutek et al. (2003) and Cesana and Antonelli (2010) considered several scenarios, such as a probability $P(W \cap R|V)$, the conditional probability that the intersection of events width and rejection occurs given that the event validity has occurred. Liu (2012) proposed a conditional probability $P(W|R)$, which is the probability of the event width conditional on the null hypothesis being rejected. Collectively, these scenarios and existing methods outline a critical role for helping practitioners and researchers appropriately accomplish their research needs and to avoid false interpretation (Ioannidis, 2005).

In view of methodological justification that has been mentioned so far, the sample size calculation methods have been mostly restricted to the cases of equal variances. In practice, there is evidence that extreme variance ratios do occur (Wilcox, 1987). While constructing CIs for two-group comparisons of means, if the homogeneity of variances cannot be assumed, the sample sizes for each group should be allocated in some way to take into account the proportionality with variances. It has already been noted that the uncertainty about population variances is very important (Grieve, 1991), and so alternative methods, such as Welch's approximate t test, have been advocated (Lee, 1992). For CI-based inferences with the case of heteroscedasticity, Wang and Kupper (1997) calculated sample sizes that were both unconditional and conditional on the coverage (i.e., inclusion of the true value). However, none of these above-mentioned works dealt with the cost of sampling.

It should be noted that formulas for sample sizes that are needed for interval estimation, especially when taking the cost factor into account, are generally not found in elementary textbooks, and are thus ignored by most researchers (Algina & Olejnik, 2000; Allison et al., 1997; Dumville et al., 2006). Hsu (1994) noted that cost ratios and unbalanced designs are seldom used to determine sample size when the total cost is limited even though the use of unequal randomization ratios offers a number of advantages. In fact, unbalanced designs are the norm rather than the exception in applied settings (Yang et al., 1996). The allocation ratios are then developed for the two-sample trimmed mean case (Guo & Luh, 2009) as well as for heterogeneous-variance group comparisons (Guo & Luh, 2013). For a fixed CI width, Pentico (1981) developed a sample size formula that minimizes the cost related to two-group estimation when variances are known. Jan and Shieh (2011) and Shieh and Jan (2012) developed optimal sample sizes of Welch's t test for testing hypotheses and for CI width, respectively. Together these studies highlight the need to have high statistical power and good precision of estimation for sample size planning so that the unconditional/conditional probability for the event width, validity, and/or rejection can be maximal if the sampling cost is constrained.

A major problem of the afore-mentioned methods suggests that developing easy-to-use computer applications for practitioners and applied researchers is critical. Thus, in the context of two-group mean comparisons where variances are unequal/unknown, the aim of the present study was to develop several R Shiny apps for practitioners to find the optimal sample size for the probability of an event of interest to fulfill two distinct motives of cost effectiveness. These motives are: (a) for achieving a desired probability level, the total sampling cost can be minimal; and (b) for a given total sampling cost, the probability of event can be maximal. In an attempt to comprehensively satisfy researchers' needs and synthesize the literature, the probabilities (cases) of event prepared in the present study contain two categories:

Unconditional: 1. $P(R)$, 2. $P(W)$, 3. $P(W \cap R)$, 4. $P(W \cap V)$, 5. $P(W \cap R \cap V)$.

Conditional: 6. $P(W|V)$, 7. $P(W \cap R|V)$, 8. $P(W|R)$, and 9. $P(W \cap V|R)$.

This thorough discussion can provide an exciting opportunity to advance our knowledge of event probability. Another advantage of our method over [Jan and Shieh \(2011\)](#) and [Shieh and Jan \(2012\)](#)'s method is that we provide simple yet accurate methods in line with those of textbook so that applied researchers can easily comprehend.

The remainder of this article is organized as follows. In the section immediately following, sample size planning while considering corresponding probabilities is proposed. In the first sub-section, the Welch test statistic is introduced and the sample size for event W is discussed. In the next sub-section, sampling cost is added on. The next section, 'Method Comparisons', describes an illustrative example and compares extant methods to assess the proposed [R Shiny apps](#). The succeeding section, 'Sample Size Tables and Simulation', presents sample size tables and a simulation study, with the last section concluding and discussing the implications.

Sample Size Planning While Considering Corresponding Probabilities

Test Statistic and Sample Sizes for Event of Interest

Let the given data X_{ij} be independent and normally distributed with mean (or expectation) μ_j , $i = 1, \dots, n_j$ (sample size), $j = 1, 2$. Suppose the two variances are σ_1^2 and σ_2^2 , and their corresponding sample variances are s_1^2 and s_2^2 , respectively, where $s_j^2 = \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 / (n_j - 1)$ and \bar{x}_j denote the sample mean of group j . Based on random samples from these populations, \bar{X}_j is normally distributed with mean (or expectation) μ_j and variance σ_j^2/n_j . To test the null hypothesis $H_0: \mu_1 - \mu_2 = 0$, we can use [Welch's \(1938\)](#) statistic, defined as

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \quad (1)$$

with approximate Type I error rate

$$P\left(\left|\frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}\right| > t_{1-\alpha/2, v} \mid H_0: \mu_1 - \mu_2 = 0\right) \leq \alpha$$

where $t_{1-\alpha/2, v}$ is the $100(1-\alpha/2)$ th percentile of the t distribution with v degrees of freedom which is calculated from the data as

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)} \quad (2)$$

(Satterthwaite, 1946; Welch, 1938). For the alternative hypothesis $H_1: \mu_1 - \mu_2 = \Delta$, the approximate power function is

$$\begin{aligned} P(R) &= P(|T'| > t_{1-\alpha/2, v} | H_1: \mu_1 - \mu_2 = \Delta) \\ &= P([T' > t_{1-\alpha/2, v}] \cup [T' < -t_{1-\alpha/2, v}] | H_1: \mu_1 - \mu_2 = \Delta) \\ &= 1 - F(t_{1-\alpha/2, v}, \nu, \delta) + F(-t_{1-\alpha/2, v}, \nu, \delta), \end{aligned}$$

where $T' = (\bar{X}_1 - \bar{X}_2) / \sqrt{s_1^2/n_1 + s_2^2/n_2}$ follows a non-central t distribution, $F(x, \nu, \delta)$ denotes the cumulative distribution function of a non-central t random variate at x with ν degrees of freedom, and the non-centrality parameter $\delta = \Delta / \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. The $100(1 - \alpha)\%$ CI, i.e., $P(V) = 1 - \alpha$, for the mean difference $\mu_1 - \mu_2 = \Delta$ is

$$\left((\bar{X}_1 - \bar{X}_2) - t_{1-\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{X}_1 - \bar{X}_2) + t_{1-\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

The CI width is then

$$2t_{1-\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (3)$$

If the CI width is less than or equal to a desired width value w that should be sensibly chosen according to the alternative hypothesis, the required group sizes can be determined by

$$n_1 \geq \left(\frac{2t_{1-\alpha/2, v}}{w} \right)^2 \left(s_1^2 + \frac{s_2^2}{k} \right) \text{ and } n_2 = kn_1 \quad (4)$$

where k is an allocation ratio.

It should be noted that, in Equation (4), s_j^2 is an estimate of σ_j^2 and is random, sometimes underestimating and sometimes overestimating σ_j^2 , $j = 1, 2$. Hence, the sample sizes in Equation (4) bring about $P(W) \approx .5$. As such, it is imperative to find adequate sample sizes such that the probability, $P(W)$, achieves the desired probability level $1 - \beta$. That is, the group sizes n_1 and n_2 must be chosen such that they satisfy

$$P(W) = \Pr \left\{ 2t_{1-\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq w \right\} \geq 1 - \beta \quad (5)$$

Based on Casella and Berger (2002),

$$\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} = H \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right) \quad (6)$$

It is known that vH is distributed approximately as χ^2 distribution with v degrees of freedom. Assuming $n_2 = kn_1$, Equation (5) can be re-written as

$$P(W) = P\{\chi^2 \leq w_1\} \geq 1 - \beta \quad (7)$$

where $\chi^2 = vH$ and $w_1 = n_1vw^2 / (4(\sigma_1^2 + \sigma_2^2/k)t_{1-\alpha/2, v}^2)$. Because Equation (7) cannot be solved explicitly, App (I) (see [Supplementary Materials](#)) was developed to find the required sample sizes to satisfy $w_1 \geq \chi_{1-\beta, v}^2$, where $\chi_{1-\beta, v}^2$ is the $100(1 - \beta)$ th percentile of the χ^2 distribution with v degrees of freedom. In addition to $P(W)$, other unconditional/conditional probabilities (cases) of event are also obtained in [Appendix](#).

Adding on Sampling Costs

In the present study, let C represent the total sampling cost, excluding any fixed overhead costs, and suppose the total cost is a linear function of sample sizes, for which the cost function is $C = n_1c_1 + n_2c_2 = n_1(c_1 + kc_2)$, where c_j is the cost of obtaining an observation from group j . If sampling cost is the concern, we need to find the optimal sample size allocation ratio, k , so that the total cost (or total sample size, $N = n_1 + n_2$) can be minimal. [Pentico \(1981\)](#) provided an optimal allocation ratio $k = (\sigma_2/\sigma_1)\sqrt{c_1/c_2}$ in a normal case when variances are known, and the resulting sample sizes reflect a minimal total cost. When variances are unknown, his method can be adopted by using some planning values of variance, such as the sample variances, to calculate the allocation ratio

$$k = (s_2/s_1)\sqrt{c_1/c_2} \quad (8)$$

For motive (a), achieving a desired probability level for minimal total cost, the provided App (I) can be executed by exhaustion algorithms ([Guo & Luh, 2020](#)), the optimal sample size (setting $k = 0$, see [Figure 1](#)) then can be obtained in any of the nine probability cases so that the total sampling cost can be minimal. To achieve optimization, researchers need to form a range of plausible values of n_1 by specifying two adjustment values, $0 < a < 1 < b$, to better execute exhaustion algorithms in [App \(I\)](#). It is suggested that $a = 0.8$ and $b = 1.2$ be used to form a narrower search space. If a warning statement “ a is too large” or “ b is too small” appears, researchers need to adjust the values to enlarge the search space. Several reminders are in order. Researchers can specify one- or two-tailed tests in the apps, and the value of the mean difference can be either positive or negative. If the researcher only considers $P(R)$, the desired width can be plugged in as any positive value. Conversely, if the event of interest is not related to rejection, then the mean difference (Δ) can be plugged in as any value. Further, setting the allocation ratio ($k = 0$) and unit costs (1, 1) can find the minimal total number of sample sizes. Note that when the allocation ratio (k) is pre-specified (e.g., 0.5, 1, 2, 3...) other than 0, [App \(I\)](#) can also be employed to find the needed sample size for any of the nine probabilities.

Figure 1

A Screenshot of R Shiny App (I)

Sample Sizes for Two Means: Test and Confidence Intervals for Cost Constraints

α , significance level	c_1 , unit cost of group 1	Δ , mean difference
<input type="text" value="0.05"/>	<input type="text" value="1"/>	<input type="text" value="5"/>
$1-\beta$, desired probability	c_2 , unit cost of group 2	Desired width of CI
<input type="text" value="0.8"/>	<input type="text" value="1"/>	<input type="text" value="7"/>
a , ($0 < a < 1$)	Planning value of σ_1^2	Case , set 1 for P(R); 2 for P(W); 3 for P(W∩R); 4 for P(W∩V); 5 for P(W∩R∩V); 6 for P(W V); 7 for P(W∩R V); 8 for P(W R); 9 for P(W∩V R);
<input type="text" value="0.8"/>	<input type="text" value="100"/>	
b , ($b > 1$)	Planning value of σ_2^2	<input type="text" value="1"/>
<input type="text" value="1.2"/>	<input type="text" value="100"/>	
Tail(s), for one or two-tailed test	k , allocation; set $k=0$ for optimal size	
<input type="text" value="2"/>	<input type="text" value="0"/>	

output

```
n1 = 64,
n2 = 64,
Minimal cost = 128,
Obtained probability = 0.80146
```

However, in this case, the desired level $1 - \beta$ can still be achieved, but the total cost may not be minimal.

For motive (b), on the other hand, if there is a budget constraint and the total sampling cost is limited, we need to choose an optimal sample size such that a maximal probability of the event can be obtained. Thus, Equation (8) for the cost function $C = n_1(c_1 + kc_2)$ can be used to obtain the initial sample size, respectively, as

$$\begin{cases} n_1 = \frac{C}{c_1 + kc_2} = \frac{s_1}{\sqrt{c_1}} \times \frac{C}{(s_1\sqrt{c_1} + s_2\sqrt{c_2})} \\ n_2 = kn_1 = \frac{s_2}{\sqrt{c_2}} \times \frac{C}{(s_1\sqrt{c_1} + s_2\sqrt{c_2})} \end{cases} \quad (9)$$

For this task, App (II) (see [Supplementary Materials](#)) is supplied. It should be noted that a floor function is used in computing to ensure that the resulting cost is no more than the given total cost C .

Method Comparisons

In this section, several comparisons are described in order to see if the proposed approach gives consistent or better results than existing methods for two-tailed test or two-sided CIs. For motive (a), firstly, when variances are equal, in [Liu's study \(2012\)](#), patients were assigned to receive either psychotherapy to reduce stress in terms of blood pressure or no intervention. The population variances for blood pressure were $\sigma_1^2 = \sigma_2^2 = 100$, the minimal effect size was $\Delta = 5$ (mmHg), and the CI width was set to be 7. We set the same condition as Table 1 of [Liu \(2012\)](#) and set the unit cost as $(c_1, c_2) = (1, 1)$. For the case of $P(R)$, we employed App (I) and found the computed sample sizes of (64, 64) with the obtained probability of .80146. For the case of $P(W|R)$, we found the computed sample sizes of (70, 70) with the obtained probability of .803865. Both cases are consistent with Liu's result. Secondly, when variances are unequal, we used the condition in Table 1 of [Lee \(1992\)](#) to set $\Delta = 1$, $\sigma_1^2 = 1.6$, $\sigma_2^2 = 0.4$, $(c_1, c_2) = (1, 1)$ for two-sided $\alpha = .05$, $\sigma_1/\sigma_2 = 2$ and $d = \Delta/\sqrt{\sigma_1^2/2 + \sigma_2^2/2} = 1$. After employing App (I), we obtained sample sizes (21, 10) for power of .8 and (27, 13) for power of .9, compared to his results of (21, 11) and (27, 14), respectively. Taken together, these results indicate consistency in terms of sample sizes. Thirdly, when unequal variances and unit costs are the concern and motive (a) is the focus, we ran all the conditions by employing App (I) as Table 4 of [Jan and Shieh \(2011\)](#) for $P(R)$; and fourthly, to consider CI width, we ran all the conditions as Table 8 of [Shieh and Jan \(2012\)](#) for $P(W)$. As a result, almost the same values of sample sizes were obtained, and only a few cases of discrepancy were identified, which are summarized in [Table 1](#) for $P(R)$ and [Table 2](#) for $P(W)$. It is apparent from these two tables that while the resulting probabilities can be satisfactorily maintained, our total cost is slightly lower than those of their methods.

For motive (b) in a fixed total cost, we employed App (II) to run all the conditions as Table 3 of [Jan and Shieh \(2011\)](#) for $P(R)$ and as Table 7 of [Shieh and Jan \(2012\)](#) for $P(W)$, respectively. Subsequently, for $P(R)$, we obtained all the same group sizes as [Jan and Shieh's \(2011\)](#) results with comparable power. For $P(W)$, we got three cases of different group sizes compared with [Shieh and Jan \(2012\)](#), which are presented in [Table 3](#).

Table 1

For $P(R)$, Computed Total Cost, Sample Sizes and Obtained Probability

(σ_1^2, σ_2^2) with (c_1, c_2)	Proposed App (I)				Jan & Shieh (2011)			
	cost	n_1	n_2	probability	cost	n_1	n_2	probability
(1,1) with (1,1)	45	22	23	.906142	45	23	22	.9057
(1,1) with (1,3)	83	29	18	.900254	84	30	18	.9032
(1/9,1) with (1,1)	21	5	16	.902258	22	6	16	.9144
(1/9,1) with (1,2)	36	6	15	.900894	37	7	15	.9086

Note. Setting $\Delta = 1$; $\alpha = .05$; $1 - \beta = .9$.

Table 2

For $P(W)$, Computed Total Cost, Sample Sizes and Obtained Probability for Motive (a)

(σ_1^2, σ_2^2) with (c_1, c_2)	Proposed App (I)				Shieh & Jan (2012)			
	cost	n_1	n_2	probability	cost	n_1	n_2	probability
(4,1) with (1,3)	253	130	41	.901892	254	128	42	.9060
(9,1) with (1,2)	337	225	56	.900217	338	226	56	.9057
(9,1) with (1,3)	390	240	50	.900709	391	238	51	.9042

Note. Setting $w = 1$; $\alpha = .05$; $1 - \beta = .9$.

Table 3

For $P(W)$, Computed Sample Sizes and Obtained Probabilities for Motive (b)

(σ_1^2, σ_2^2) with (c_1, c_2)	Fixed total	Proposed App (II)			Shieh & Jan (2012)		
	cost	n_1	n_2	probability	n_1	n_2	probability
(1/9,1) with (1,3)	50	8	14	.152415	11	13	.1546
(1,1) with (1,3)	80	35	15	.066037	38	14	.0679
(4,1) with (1,2)	180	104	38	.470735	106	37	.4723

Note. Setting $w = 1$; $\alpha = .05$.

Sample Size Tables and Simulation

It should be noted that except for the true mean difference and the desired width, the values of variances and unit costs are also the key elements for allocating group sizes. Thus, these four parameters were varied to present the features of sample size planning in the following two sample size tables for the nine probability cases (all sample sizes are rounded up to the nearest integer) for motive (a). In Table 4 for equal variances/unit costs, there are six configurations of (Δ, w) , arranged with a small (or large) effect together with a narrow (or wide) desired width, while in Table 5 there are four configurations for unequal variances/unit costs. These tables are quite revealing in several ways. First, the impacts of the true mean difference and the desired width can be unfolded in Table 4 by taking Case 1 ($P(R)$) and Case 2 ($P(W)$) as cross-references. In Case 1, as the true value Δ increases from 2 to 8 (see Columns 1-3), the sample sizes decrease from (393, 394) to (26, 26), showing that it is not influenced by the value of w at all.

Table 4

Optimal Sample Sizes (n_1, n_2) for the Nine Cases of Probabilities

Case	(Δ, w)		
	(2, 3)	(4, 3)	(8, 3)
1. $P(R)$	393, 394	99, 100	26, 26
2. $P(W)$	358, 358	358, 358	358, 358
3. $P(W \cap R)$	395, 395	358, 358	358, 358
4. $P(W \cap V)$	361, 361	361, 361	361, 361
5. $P(W \cap R \cap V)$	420, 420	361, 361	361, 361
6. $P(W V)$	358, 358	358, 358	358, 358
7. $P(W \cap R V)$	385, 386	358, 358	358, 358
8. $P(W R)$	357, 358	358, 358	358, 358
9. $P(W \cap V R)$	359, 360	361, 361	361, 361
Case	(2, 4)	(4, 4)	(8, 4)
1. $P(R)$	393, 394	99, 100	26, 26
2. $P(W)$	205, 205	205, 205	205, 205
3. $P(W \cap R)$	393, 394	205, 206	205, 205
4. $P(W \cap V)$	207, 207	207, 207	207, 207
5. $P(W \cap R \cap V)$	420, 420	207, 207	207, 207
6. $P(W V)$	205, 205	205, 205	205, 205
7. $P(W \cap R V)$	379, 379	205, 205	205, 205
8. $P(W R)$	204, 204	204, 205	205, 205
9. $P(W \cap V R)$	206, 206	206, 206	207, 207

Note. Setting $\alpha = .05$; $1 - \beta = .8$; $\sigma_1 = \sigma_2 = 10$; unit costs (1,1).

Table 5

Optimal Sample Sizes (n_1, n_2) for the Nine Cases of Probabilities Given Various Values of σ_2 and Unit Costs (c_1, c_2)

Case	(1)	(2)	(3)	(4)
	$\sigma_2 = 10;$ $(c_1, c_2) = (1, 1)$	$\sigma_2 = 5;$ $(c_1, c_2) = (1, 1)$	$\sigma_2 = 5;$ $(c_1, c_2) = (1, 4)$	$\sigma_2 = 5;$ $(c_1, c_2) = (4, 1)$
1. $P(R)$	64, 64	49, 24	63, 17	41, 39
2. $P(W)$	36, 37	29, 15	36, 11	25, 23
3. $P(W \cap R)$	64, 64	49, 24	63, 17	41, 39
4. $P(W \cap V)$	37, 38	30, 15	39, 11	26, 23
5. $P(W \cap R \cap V)$	68, 69	52, 26	68, 18	44, 41
6. $P(W V)$	36, 37	29, 15	37, 11	26, 20
7. $P(W \cap R V)$	61, 62	47, 24	62, 16	40, 36
8. $P(W R)$	36, 36	28, 14	37, 10	24, 23
9. $P(W \cap V R)$	36, 37	29, 15	36, 11	25, 23

Note. Setting $\alpha = .05; 1 - \beta = .8; \Delta = 5; w = 10; \sigma_1 = 10$.

These sample sizes are consistent with the results in Table 2 of Liu’s study (2012). In Case 2, on the contrary, as the desired value w increased from 3 to 4 (see the upper and lower panels), the needed sample sizes dramatically reduced, regardless of the value of Δ . The fact of the independence of Δ can also be observed in Cases 4 and 6. Further note that observing the sample sizes of Cases 1 and 2 shows that many discrepancies were clearly demonstrated; that is, the sample sizes for hypothesis testing and constructing CIs are different, which was mentioned in Borenstein et al. (2001). Second, the more events that are considered, the greater the required sample sizes. It can be seen that among Cases 1-5, the relatively greater sample size required appears in Case 5 which is the intersection of events $W, R,$ and V . Third, if the probability is conditional on the event validity, the required sample sizes are relatively fewer than those of intersection events (see Case 6 vs. Case 4 and Case 7 vs. Case 5). Note that $P(W|V) = P(W \cap V)/P(V) = P(W \cap V)/(1 - \alpha)$ and thus $P(W|V) \geq P(W \cap V)$. Similarly, if the probability is conditional on the event rejection, the required sample sizes are also fewer than those of intersection events (see Case 8 vs. Case 3 and Case 9 vs. Case 5). Finally, if $\Delta/w < 1$, the resulting sample sizes of Cases 2-9 show greater discrepancies than those cases of $\Delta/w \geq 1$. The smaller the value of Δ/w , the greater the discrepancy. It can be seen that the discrepancy of sample sizes for $(\Delta, w) = (2, 4)$ are greater than those for $(\Delta, w) = (2, 3)$. To sum up, the various sample sizes among the nine cases significantly emerge under different parameter configurations and events of interest.

In Table 5, we considered unequal variances/unit costs and demonstrate four configurations. Take Column 1 as a reference, where variances and unit costs are both equal.

It can be seen from Column 2 that approximately half the sample size is needed for the second group because the variance in that group is reduced. In other words, the allocation k is 0.5. Then, take Column 2 as a reference where unit costs are equal. As can be seen from Column 3, a much smaller sample size is required for the second group due to the higher unit cost in that group. In this situation, the allocation k is 0.25. Finally, Column 4 shows approximately equal group sizes because larger variance and higher cost are both taken into account. Note that although the allocation k is 1 in this situation, the resulting group sizes may not be equal because the algorithms in the apps seek optimization.

To carry out a computer simulation, we chose the optimal sample sizes (26, 20) in Case 6 $P(W|V)$ from Table 5, given $\Delta = 5$, $w = 10$ and ran 10,000-replicated simulations. The results show that the average mean difference is 5.0049 ($SD = 2.2633$) and the coverage rate is 0.9503, indicating good coverage. The average width for which CIs cover the true mean difference is 9.1116, while the average width for which CIs do not cover the true mean difference is 8.5862. Accordingly, the empirical probabilities of the nine cases are 1. $P(R) = .5787$, 2. $P(W) = .8100$, 3. $P(W \cap R) = .4951$, 4. $P(W \cap V) = .7637$, 5. $P(W \cap R \cap V) = .4709$, 6. $P(W|V) = .8036$, 7. $P(W \cap R|V) = .4955$, 8. $P(W|R) = .8555$, and 9. $P(W \cap V|R) = .8137$. As expected, the resulting probability of Case 6 is very close to the nominal level of .8, and Cases 1, 3, 4, 5, and 7 cannot achieve the nominal level due to the given sample size (26, 20) being insufficient (see Column 4 of Table 5).

Conclusions and Discussion

The current discussions in sample size planning is to fulfill one or more goals, such as power-based statistical tests, precision of estimations, minimal costs, or some other criteria. To avoid the pitfall of treating inferential statistics as descriptive statistics, the probabilistic thinking is essential for the scheme of sample size determination. Along with rapid advances of ready-to-used computer applications, the present study aimed to contribute to this growing area and developed R Shiny App (I) to find sample sizes to minimize a sampling total cost for a desired probability level and App (II) to maximize the probability of an event of interest for a fixed total cost. The choice of nine probability cases of events can be selected in the apps, depending on the purpose of the research. The present study also showed that the sample size allocation ratio is affected by the proportion of standard deviations and the disproportion of the square roots of the unit costs. The idea is to obtain more sample sizes for the group that has larger variance or for the group with the lower unit cost. When the unit costs or the standard deviations are dissimilar across groups, the use of optimal sample sizes can result in substantial savings. The interrelation among the values of true difference, the desired width, the variances, and the unit costs as well as the events of interest all affect the seeking of optimal sample

sizes. Their roles are just like a five-way tug of war. This study provides an important opportunity to advance the understanding of sample size planning.

As [Lenth \(2001\)](#) noted, not all sample size problems are the same. [Cohen \(1990\)](#) stated that “I have learned that there is no royal road to statistical induction, that the informed judgment of the investigator is the crucial element in the interpretation of data, and that things take time.” Among other things, a desired width can be sensibly chosen according to the alternative hypothesis ([Cesana & Antonelli, 2010](#)) and viable benchmarks for the width among disciplines may be established for improving meta-analysis ([Smithson, 2003](#)). Since a larger sample size is needed for a smaller $|\Delta|$ and/or w , one of the issues that emerges from the present findings is to sort out the relation between the mean difference and a desired width. An implication of this is the possibility to consider a suitable range of alternative true mean difference as $0.2\sigma \leq |\Delta| \leq 0.8\sigma$, setting $\sigma = \sqrt{\sigma_1\sigma_2}$ (geometric mean), i.e., from a small to large effect based on Cohen’s effect size paradigm. Moreover, a suitable range of desired width is suggested as $0.3\sigma \leq w \leq \min(2|\Delta|, 1.4\sigma)$, which will result in a reasonable sample size value from about 20 to 400.

Since the advocacy of CIs is a central theme in statistical reform, additional methods and formulas for sample size determination in regard to various unconditional/conditional probabilities are required to address this need in the future to simultaneously obtain power and precision. The current findings add to a growing body of literature on sample size planning, and there is abundant room for further progress in multiple comparisons (simultaneous confidence intervals) ([Liu, 2009](#); [Pan & Kupper, 1999](#)), linear contrasts ([Bonett, 2009](#); [Jan & Shieh, 2019](#); [Luh & Guo, 2016](#)), stratified sampling ([Snedecor & Cochran, 1989](#)), and multilevel analyses ([Liu, 2003](#)). Under certain circumstances, one group size may be pre-fixed to a certain number in a study; thus, the present study also supplied App (III) (see [Supplementary Materials](#)) to facilitate the recommended procedures in planning the required size of the other group so that the desired probability level can still be achieved.

Funding: This research was supported by a National Science Council grant, Taiwan (NSC98-2410-H-006-067-MY3).

Acknowledgments: The author thanks Emeritus Professor Jiin-Huang Guo of National Pingtung University, Taiwan for his guidance on derivation and programming.

Competing Interests: The author has declared that no competing interests exist.

Supplementary Materials

For this article, three R Shiny apps can be found online to facilitate integrative consideration on sample size planning (for access see [Index of Supplementary Materials](#) below):

- App I: Test and Confidence Intervals for Cost Constraints

- App II: Test and Confidence Intervals for Fixed Total Cost
- App III: Test and Confidence Intervals for Fixed n_1

Index of Supplementary Materials

- Luh, W. M. (2022). *Supplementary materials to "Probabilistic thinking is the name of the game: Integrating test and confidence intervals to plan sample sizes". Sample Sizes for Two Means: Test and Confidence Intervals for Cost Constraints* [R Shiny App].
<https://sample-size-ci.shinyapps.io/R-Size-Mean-CI/>
- Luh, W. M. (2022). *Supplementary materials to "Probabilistic thinking is the name of the game: Integrating test and confidence intervals to plan sample sizes". Sample Sizes for Two Means: Test and Confidence Intervals for Fixed Total Cost* [R Shiny App].
<https://sample-size-ci.shinyapps.io/Size-Mean-Fix-Cost/>
- Luh, W. M. (2022). *Supplementary materials to "Probabilistic thinking is the name of the game: Integrating test and confidence intervals to plan sample sizes". Sample Sizes for Two Means: Test and Confidence Intervals for Fixed n_1* [R Shiny App].
<https://sample-size-ci.shinyapps.io/Size-Mean-Fix-N1/>

References

- Algina, J., & Olejnik, S. (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research*, *35*(1), 119–137.
https://doi.org/10.1207/S15327906MBR3501_5
- Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods*, *2*(1), 20–33. <https://doi.org/10.1037/1082-989X.2.1.20>
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.).
- Beal, S. L. (1989). Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics*, *45*(3), 969–977.
<https://doi.org/10.2307/2531696>
- Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*(4), 389–396.
<https://doi.org/10.1037/1082-989X.10.4.389>
- Bonett, D. G. (2009). Estimating standardized linear contrasts of means with desired precision. *Psychological Methods*, *14*(1), 1–5. <https://doi.org/10.1037/a0014270>
- Borenstein, M., Rothstein, H., & Cohen, J. (2001). *Power and precision*. Biostat.
- Calin-Jageman, R. J., & Cumming, G. (2019). The new statistics for better science: Ask how much, how uncertain, and what else is known. *The American Statistician*, *73*(Suppl. 1), S271–S280.
<https://doi.org/10.1080/00031305.2018.1518266>
- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury/Thomson Learning.

- Cesana, B. M., & Antonelli, P. (2010). A new approach to sample size calculations for the power of testing and estimating population means of Gaussian distributed variables. *Biomedical Statistics and Clinical Epidemiology*, *4*(2), 67–78.
- Cohen, J. (1990). Things I have learned (so far). *The American Psychologist*, *45*(12), 1304–1312. <https://doi.org/10.1037/0003-066X.45.12.1304>
- Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics: Estimation, open science, and beyond*. Taylor & Francis.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*(4), 532–574. <https://doi.org/10.1177/0013164401614002>
- Dumville, J. C., Hahn, S., Miles, J. N. V., & Torgerson, D. J. (2006). The use of unequal randomization ratios in clinical trials: A review. *Contemporary Clinical Trials*, *27*(1), 1–12. <https://doi.org/10.1016/j.cct.2005.08.003>
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*(2), 119–126. <https://doi.org/10.1111/j.0963-7214.2004.01502008.x>
- Grieve, A. P. (1989). Letter to the Editor. *Lancet*, *1*, 337.
- Grieve, A. P. (1991). Confidence intervals and sample sizes. *Biometrics*, *47*(4), 1597–1603. <https://doi.org/10.2307/2532411>
- Guo, J. H., & Luh, W. M. (2009). Optimum sample size allocation to minimize cost or maximize power for the two-sample trimmed mean test. *British Journal of Mathematical & Statistical Psychology*, *62*(2), 283–298. <https://doi.org/10.1348/000711007X267289>
- Guo, J. H., & Luh, W. M. (2013). Efficient sample size allocation with cost constraints for heterogeneous-variance group comparison. *Journal of Applied Statistics*, *40*(12), 2549–2563. <https://doi.org/10.1080/02664763.2013.819417>
- Guo, J. H., & Luh, W. M. (2020). Testing two variances for superiority/non-inferiority and equivalence: Using the exhaustion algorithm for sample size allocation with cost. *British Journal of Mathematical & Statistical Psychology*, *73*(2), 316–332. <https://doi.org/10.1111/bmsp.12172>
- Hsu, L. M. (1994). Unbalanced designs to maximize statistical power in psychotherapy efficacy studies. *Psychotherapy Research*, *4*(2), 95–106. <https://doi.org/10.1080/10503309412331333932>
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), Article e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Jan, S.-L., & Shieh, G. (2011). Optimal sample sizes for Welch's test under various allocation and cost considerations. *Behavior Research Methods*, *43*, 1014–1022. <https://doi.org/10.3758/s13428-011-0095-7>
- Jan, S.-L., & Shieh, G. (2019). Optimal contrast analysis with heterogeneous variances and budget concerns. *PLoS One*, *14*(3), Article e0214391. <https://doi.org/10.1371/journal.pone.0214391>

- Jiroutek, M. R., Muller, K. E., Kupper, L. L., & Stewart, P. W. (2003). A new method for choosing sample size for confidence interval-based inferences. *Biometrics*, *59*(3), 580–590.
<https://doi.org/10.1111/1541-0420.00068>
- Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation & the Health Professions*, *26*(3), 258–287. <https://doi.org/10.1177/0163278703255242>
- Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363–385. <https://doi.org/10.1037/1082-989X.11.4.363>
- Lai, K., & Kelley, K. (2012). Accuracy in parameter estimation for ANCOVA and ANOVA contrasts: Sample size planning via narrow confidence intervals. *British Journal of Mathematical & Statistical Psychology*, *65*(2), 350–370. <https://doi.org/10.1111/j.2044-8317.2011.02029.x>
- Lee, A. F. S. (1992). Optimal sample sizes determined by two-sample Welch's *t* test. *Communications in Statistics - Simulation and Computation*, *21*(3), 689–696.
<https://doi.org/10.1080/03610919208813043>
- Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*(3), 187–193. <https://doi.org/10.1198/000313001317098149>
- Liu, X. (2003). Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomization. *Journal of Educational and Behavioral Statistics*, *28*(3), 231–248. <https://doi.org/10.3102/10769986028003231>
- Liu, X. S. (2009). Sample size and the width of the confidence interval for mean difference. *British Journal of Mathematical & Statistical Psychology*, *62*(2), 201–215.
<https://doi.org/10.1348/000711008X276774>
- Liu, X. S. (2012). Implications of statistical power for confidence intervals. *British Journal of Mathematical & Statistical Psychology*, *65*(3), 427–437.
<https://doi.org/10.1111/j.2044-8317.2011.02035.x>
- Luh, W. M., & Guo, J. H. (2016). Sample size planning for the noninferiority or equivalence of a linear contrast with cost considerations. *Psychological Methods*, *21*(1), 13–34.
<https://doi.org/10.1037/met0000039>
- Meeker, W. Q., Hahn, G. J., & Escobar, L. A. (2017). *Statistical intervals: A guide for practitioners and researchers* (2nd ed.). Wiley.
- Mendoza, J. L., & Stafford, K. L. (2001). Confidence intervals, power calculation, and sample size estimation for the squared multiple correlation coefficient under the fixed and random regression models: A computer program and useful standard tables. *Educational and Psychological Measurement*, *61*(4), 650–667. <https://doi.org/10.1177/00131640121971419>
- Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123.
<https://doi.org/10.3758/s13423-015-0947-8>

- Pan, Z., & Kupper, L. L. (1999). Sample size determination for multiple comparison studies treating confidence interval width as random. *Statistics in Medicine*, *18*(12), 1475–1488.
[https://doi.org/10.1002/\(SICI\)1097-0258\(19990630\)18:12<1475::AID-SIM144>3.0.CO;2-0](https://doi.org/10.1002/(SICI)1097-0258(19990630)18:12<1475::AID-SIM144>3.0.CO;2-0)
- Pentico, D. W. (1981). On the determination and use of optimal sample sizes for estimating the difference in means. *The American Statistician*, *35*(1), 40–42.
<https://doi.org/10.1080/00031305.1981.10479301>
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110–114. <https://doi.org/10.2307/3002019>
- Shieh, G., & Jan, S. L. (2012). Optimal sample sizes for precise interval estimation of Welch's procedure under various allocation and cost considerations. *Behavior Research Methods*, *44*, 202–212. <https://doi.org/10.3758/s13428-011-0139-z>
- Smithson, M. (2000). *Statistics with confidence*. SAGE.
- Smithson, M. (2003). *Confidence intervals*. SAGE.
- Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Iowa State University Press.
- Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *The American Psychologist*, *53*(7), 799–800. <https://doi.org/10.1037/0003-066X.53.7.799>
- Trafimow, D. (2018). Confidence intervals, precision and confounding. *New Ideas in Psychology*, *50*, 48–53. <https://doi.org/10.1016/j.newideapsych.2018.04.005>
- Trafimow, D., & Uhalt, J. (2020). The inaccuracy of sample-based confidence intervals to estimate a priori ones. *Methodology*, *16*(2), 112–126. <https://doi.org/10.5964/meth.2807>
- Wang, Y., & Kupper, L. L. (1997). Optimal sample sizes for estimating the difference in means between two normal populations treating confidence interval length as a random variable. *Communications in Statistics - Theory and Methods*, *26*(3), 727–741.
<https://doi.org/10.1080/03610929708831945>
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (Eds.). (2019). Moving to a world beyond "p > 0.05" *The American Statistician*, *73*(Suppl.1), 1–401.
- Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*(3/4), 350–362. <https://doi.org/10.1093/biomet/29.3-4.350>
- Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology*, *38*, 29–60.
<https://doi.org/10.1146/annurev.ps.38.020187.000333>
- Yang, H., Sackett, P. R., & Arvey, R. D. (1996). Statistical power and cost in training evaluation: Some new considerations. *Personnel Psychology*, *49*(3), 651–668.
<https://doi.org/10.1111/j.1744-6570.1996.tb01588.x>

Appendix

To prepare the probability of the intersection event of W and R , $P(W \cap R)$, from Equation (6) in the present study and Liu (2012), the power function $P(R)$ can be expressed as

$$\begin{aligned}
P(R) &= P(|T'| > t_{1-\alpha/2, v} | H_1: \mu_1 - \mu_2 = \Delta) \\
&= P\left(\left[\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} > t_{1-\alpha/2, v}\right] \cup \left[\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} < -t_{1-\alpha/2, v}\right] \middle| H_1: \mu_1 - \mu_2 = \Delta\right) \\
&= P\left(\left[\frac{\bar{X}_1 - \bar{X}_2 - \Delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} > \frac{\sqrt{s_1^2/n_1 + s_2^2/n_2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} t_{1-\alpha/2, v} - \frac{\Delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right] \cup \right. \\
&\quad \left. \left[\frac{\bar{X}_1 - \bar{X}_2 - \Delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < -\frac{\sqrt{s_1^2/n_1 + s_2^2/n_2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} t_{1-\alpha/2, v} - \frac{\Delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right] \middle| H_1: \mu_1 - \mu_2 = \Delta\right) \\
&= P([Z > t_{1-\alpha/2, v} \sqrt{\chi^2/v} - \delta] \cup [Z < -t_{1-\alpha/2, v} \sqrt{\chi^2/v} - \delta] | H_1: \mu_1 - \mu_2 = \Delta) \\
&= P([Z > q - \delta] \cup [Z < -q - \delta] | H_1: \mu_1 - \mu_2 = \Delta)
\end{aligned}$$

where Z is a standard normal random variable, χ^2 is a chi-squared random variable with v degrees of freedom as Equation (2), $\delta = \Delta / \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ is a non-centrality parameter, and $q = t_{1-\alpha/2, v} \sqrt{\chi^2/v}$. Hence,

$$\begin{aligned}
P(W \cap R) &= P\{\chi^2 \leq w_1\} \cap ([Z > q - \delta] \cup [Z < -q - \delta]) \\
&= P\{(\chi^2 \leq w_1) \cap [Z > q - \delta]\} \cup \{(\chi^2 \leq w_1) \cap [Z < -q - \delta]\}
\end{aligned}$$

For $P(W \cap V)$, first, $P(V)$ can be expressed as

$$\begin{aligned}
P(V) &= P\left((\bar{X}_1 - \bar{X}_2) - t_{1-\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \Delta < (\bar{X}_1 - \bar{X}_2) + t_{1-\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) \\
&= P\left(-\frac{\sqrt{s_1^2/n_1 + s_2^2/n_2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} t_{1-\alpha/2, v} < \frac{\bar{X}_1 - \bar{X}_2 - \Delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < \frac{\sqrt{s_1^2/n_1 + s_2^2/n_2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} t_{1-\alpha/2, v}\right) \\
&= P(-t_{1-\alpha/2, v} \sqrt{\chi^2/v} < Z < t_{1-\alpha/2, v} \sqrt{\chi^2/v}) = P(-q < Z < q)
\end{aligned}$$

Then,

$$\begin{aligned}
P(W \cap V) &= P\{\chi^2 \leq w_1\} \cap [-t_{1-\alpha/2, v} \sqrt{\chi^2/v} < Z < t_{1-\alpha/2, v} \sqrt{\chi^2/v}] \\
&= P\{\chi^2 \leq w_1\} \cap [-q < Z < q]
\end{aligned}$$

For the probability of the intersection of events W , R , and V ,

$$\begin{aligned}
P(W \cap R \cap V) &= P\{\chi^2 \leq w_1\} \cap ([Z > q - \delta] \cup [Z < -q - \delta]) \cap [-q < Z < q] \\
&= P\{\chi^2 \leq w_1\} \cap ([\max(q - \delta, -q) < Z < q] \\
&\quad \cup [-q < Z < \min(-q - \delta, q)])
\end{aligned}$$

For $\delta > 0$, $\min(-q - \delta, q) = -q - \delta < -q$, and the set $[-q < Z < \min(-q - \delta, q)]$ is empty. Thus, $P(W \cap R \cap V) = P\{\chi^2 \leq w_1\} \cap [\max(q - \delta, -q) < Z < q]$. For $\delta < 0$,

$\max(q - \delta, -q) = q - \delta > q$, and the set $[\max(q - \delta, -q) < Z < q]$ is empty. Hence,
 $P(W \cap R \cap V) = P\{\chi^2 \leq w_1 \cap [-q < Z < \min(-q - \delta, q)]\}$.

Finally, other conditional probabilities can be found as follows: $P(W|V) = P(W \cap V)/(1 - \alpha)$,
 $P(W \cap R|V) = P(W \cap R \cap V)/(1 - \alpha)$, $P(W|R) = P(W \cap R)/P(R)$, $P(W \cap V|R) = P(W \cap R \cap V)/P(R)$.



Methodology is the official journal
of the European Association of
Methodology (EAM).



leibniz-psychology.org

PsychOpen GOLD is a publishing
service by Leibniz Institute for
Psychology (ZPID), Germany.