Ahi Evran Üniversitesi
Kırşehir Eğitim Fakültesi Dergisi

ISSN: 2147 - 1037

http://kefad.ahievran.edu.tr

Statistical Power Analysis with pwrss R Package

Metin Buluş
Cahit Polat

Article Information

DOI: 10.29299/kefad.1209913

Received: 25.11.2022
Revised: 07.03.2023
Accepted: 11.04.2023

Keywords:

Statistical Power Analysis,

Minimum Required Sample Size,

pwrss R Package

Abstract

This study presents the theoretical foundations and computational approaches to statistical power analysis. Ten hypothesis tests and their derivatives are reviewed, including the test of one proportion against a constant, difference between two proportions, one mean against a constant, difference between two means (independent and matched samples), one correlation against a constant, difference between two correlations, R-squared deviation from zero in linear regression, difference between two R-squared values in hierarchical linear regression, analyses of variance/covariance (one-way, two-way and three-way ANOVA or ANCOVA) for comparing means of two or more groups, and repeated measures ANOVA. The concept of statistical power and sample size calculations for these tests are consolidated with practical examples. The hypothesis tests of non-inferiority, superiority, and equivalence, which are widely used in medical and pharmaceutical research are also introduced, and their applications are demonstrated using examples from behavioral and educational research. Calculations were performed with the pwrss R package (https://pwrss.shinyapps.io/lang-en/).

pwrss R Paketi ile İstatistiksel Güç Analizi

Makale Bilgileri

DOI: 10.29299/kefad.1209913

Yükleme: 25.11.2022
Düzeltme: 07.03.2023
Kabul: 11.04.2023

Anahtar Kelimeler:

İstatistiksel Güç Analizi,

Örneklem Büyüklüğü,

pwrss R Paketi

Öz

Bu çalışmada, yaygın olarak kullanılan hipotez testleri ışığında istatistiksel güç analizinin teorik altyapısı ve hesaplama yaklaşımları ele alınmıştır. On adet hipotez testi ve türevleri incelenmiştir; tek bir oranın bir sabite karşı, iki oranın farkı, tek bir ortalamanın bir sabite karşı, iki ortalamanın farkı (bağımlı ve bağımsız örneklemler), tek bir korelasyonun sıfırdan farkı, iki korelasyon farkı, doğrusal regresyondaki R-kare değerinin sıfırdan farkı, hiyerarşik doğrusal regresyonda iki R-kare farkı, iki ya da daha fazla grup ortalamalarının karşılaştırılması (tek faktörlü, iki faktörlü ve üç faktörlü ANOVA ya da ANCOVA) ve tekrarlı ölçümler ANOVA. Bu testler için istatistiksel güç kavramı ve örneklem büyüklüğü hesaplamaları uygulamalı örnekler ile pekiştirilmiştir. Ayrıca, tıbbi ve farmasötik araştırmalarda yaygın olarak kullanılan non-inferiority (pratik anlamda eşit ya da üstün olma), superiority (pratik anlamda üstün olma), ve equivalence (pratik anlamda eşit olma) hipotez testleri de tanıtılmış olup davranış ve eğitim bilimleri araştırmalarından örnek uygulamalar gösterilmiştir. Hesaplamalar, pwrss R paketi kullanılarak gerçekleştirilmiştir (https://pwrss.shinyapps.io/lang-tr/).

Sorumlu Yazar: Metin Buluş, Doç. Dr., Adıyaman Üniversitesi, Türkiye, bulusmetin@gmail.com, ORCID ID: 0000-0003-4348-6322.

Yazar2: Cahit Polat, Dr. Öğr. Üyesi, Harran Üniversitesi, Türkiye, cahitpolat@harran.edu.tr, ORCID ID: 0000-0002-1423-5084.

Alt Bilgi: Bu çalışma daha önce VIII. Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde özet bildiri olarak sunulmuştur.

Atıf için: Buluş, M., & Polat, C. (2023). pwrss R paketi ile istatistiksel güç analizi. *Kırşehir Eğitim Fakültesi Dergisi*, *24*(3), 2207 – 2328.

Giriş

İncelemek istenilen tüm birimleri kapsayan küme *kitle* olarak adlandırılır. Kaynaklar sınırlı olduğu için kitleyi oluşturan tüm birimlere ulaşmak çoğu zaman neredeyse imkânsızdır. Bundan dolayı, kitle karakteristiklerini barındırdığı ve kitleyi temsil ettiği düşünülen bir alt kümeden veri toplanır ve analiz edilir. Kitleden daha küçük, ekonomik ve yönetilebilir bu temsili alt küme *örneklem* olarak adlandırılır. İstatistiğin temel problemlerinden biri, doğru çıkarımlarda bulunmak için örneklem büyüklüğünün en az ne kadar olması gerektiği sorusudur. Bu sorunu derinlemesine incelemeden önce istatistikteki bazı temel kavramları açıklamak gerekmektedir.

Herhangi bir özelliğin (değişkenin ya da değişkenler arasındaki ilişkinin) kitle değeri *parametre*, kitleyi temsil ettiği düşünülen örneklemdeki değeri ise *istatistik* olarak adlandırılır. Örneklemden elde edilen istatistik kitle parametresinin bir tahmin edicisidir. Kitledeki tüm birimler kullanılmadığından, istatistik ile kitle parametresinin aynı olacağı öne sürülmez fakat bu iki değerin birbirlerine yakın olması beklenir. Daha da ötesi, yeni bir örneklem seçildiğinde elde edilen istatistik değeri de farklı olabilmekte ve örneklemden örnekleme değişebilmektedir. Kitle parametresinden meydana gelen bu sapmalar örneklem temsilinden kaynaklanan hatalardır ve istatistiksel modellerin doğru tanımlandığı durumlarda *istatistiğin standart hatası* şeklinde ifade edilir.

Bilimsel çalışmalarda örneklemden elde edilen istatistik ile birlikte istatistiğin standart hatası da raporlanır. Standart hata hakkında bilgi mevcut ise, standart hatayı makul düzeyde tutacak bir örneklem büyüklüğü çalışma öncesinde belirlenebilir. Sonsuz sayıda birimin olduğu bir örneklem elde etmek imkânsızdır. Ayrıca, tek bir birimden oluşan ya da tek bir gözlemin yapıldığı örneklem de kabul edilemez. Bir niteliğin değişken olarak kabul edilebilmesi ve istatistiğin hesaplanabilmesi için en az iki farklı gözlemin gerçekleşmiş olması gerekmektedir. İki uç değer arasındaki standart hatanın makul düzeyde olabileceği bir örneklem büyüklüğü vardır.

Gereğinden küçük örneklem pratikte önemli olan etkilerin (ya da farkların) tespit edilmesini zorlaştırır. Bundan dolayı kullanılan kaynaklar boşa gider ve katılımcılar gereksiz risk almış olur. Gereğinden büyük örneklemde ise pratikte önemli olmayacak etkileri bulmak için gereğinden fazla kaynakların kullanılması söz konusudur ve gereğinden fazla katılımcı risk almış olur. Bahsedilen etik ve ekonomik nedenlerden dolayı, standart hatayı makul düzeyde tutacak örneklem büyüklüğünü belirlemek gerekir. Çeşitli kıstaslar göz önünde bulundurularak araştırmalar için gerekli en küçük örneklem büyüklüğü *istatistiksel güç analizi* ile belirlenir.

Bilimsel çalışmaların raporlanması için oluşturulan uluslararası standartlarda istatistiksel güç analizinin yapılması önem arz etmektedir (örn. *What Works Clearinghouse*, *Strengthening the Reporting of Observational studies in Epidemiology*, *Consolidated Standards of Reporting Trials*). Literatürde güç analizi konusunda çok sayıda kaynak bulunmasına rağmen (örn., Aberson, 2019; Cohen, 1988; Hedberg, 2017; Liu, 2013; Myors ve diğerleri, 2023; Zhang ve Yuan, 2018), Türkiye'de özellikle sosyal ve beşeri bilimleri

alanında bu konuya yeterli önemin verilmediği görülmektedir. Türkiye'de 2010-2020 yılları arasında eğitim ve psikoloji bilimleri alanında raporlanan deneysel çalışmaların temsili bir örneklemini inceleyen Bulus ve Koyuncu (2021), 155 deneysel çalışmadan hiçbirinin örneklem büyüklüğünü belirlemek için güç analizi hesaplamalarına yer vermediklerini tespit etmişlerdir. Benzer şekilde, Şevgin ve Çetin (2017) Türkiye'de eğitim bilimleri alanındaki dergilerden üç tanesini rastgele seçmiş, bu dergilerde 2014 ve 2016 yılları arasında yayınlanmış 25 adet nicel çalışmayı incelemiş ve sonuç olarak hiçbirinin güç analiz yapmadıklarını ortaya koymuşlardır.

Bu durum, araştırmacıların güç analizi konusunda bilgilendirilmesinin önemini ve konuyla ilgili ulaşılabilir kapsamlı kaynakların gerekliliğini göstermektedir. Son zamanlarda Türkiye'de özellikle biyoistatistik alanında açık erişim güç analizi hesaplama araçları konusunda girişimler olsa da (örn. Arslan ve diğerleri., 2018) bu çabaların eğitim ve davranış bilimlerindeki araştırmalara yansımadığı görülmektedir. Bundan dolayı, bu çalışmanın amacı yaygın olarak kullanılan hipotez testleri ışığında istatistiksel güç analizinin teorik altyapısı ve hesaplama yaklaşımını açıklamaktır. Çalışmada hipotez testleri ve türleri için güç analizi ve örneklem büyüklüğü belirleme işlemleri uygulamalı bir şekilde eğitim ve davranış bilimleri alanından örneklerle pekiştirilmiştir.

Güç Analizinde Göz Önünde Bulundurulması Gereken Parametreler

Güç analizi ile standart hatayı makul düzeyde tutacak örneklem büyüklüğünü belirlemek için objektif bazı ölçütlere ihtiyaç vardır. Bu objektif ölçütlerin belirlenmesinde göz önünde bulundurulması gereken bazı noktalar vardır. Bunlar tip I hata, tip II hata, hipotez testinin yönü, pratik anlamda en küçük anlamlı etki, pratik anlamda ihmal edilebilecek sınır değer ve hipotez testinin tipi olarak listelenebilir.

Tip I ve Tip II Hata

Gerçekte kitle parametresi bilenmediği için alternatif hipotez (H_A) doğru olabileceği gibi yokluk hipotezinin (H_0) de doğru olma ihtimali vardır. Hangisinin gerçekte doğru olduğuna bağlı olarak yapılan çıkarsamalarda yanılmalar söz konusu olabilir. Yokluk hipotezi kitlede doğru iken örneklemde reddedebilir ya da alternatif hipotez kitlede doğru iken örneklemde yokluk hipotezi reddedilemeyebilir. Gerçekte kitlede olmayan bir etkinin (H_0 doğru, H_A yanlış) örneklemde var olduğu çıkarımını yapmak, yani H_0 hipotezini yanlışlıkla reddetmek *tip I hata* (α) olarak adlandırılır. Çoklu karşılaştırmaların ve çoklu çıktı değişkenlerinin olmadığı çalışmalarda genelde $\alpha = 0.05$ olarak belirlenir. Bu değer, örneğin hipotetik olarak 100 örneklem seçildiğinde, bunlardan en fazla 5 tanesinde tip I hata yapmanın göze alınabildiği anlamına gelir.

Gerçekte kitlede var olan bir etkinin (H_0 yanlış, H_A doğru) örneklemde yoktur çıkarımı yapmak, yani H_0 hipotezini yanlışlıkla reddedememek ise *tip II hata* (β) olarak adlandırılır. Genelde $\beta = 0.20$ olarak tanımlanır ve bu değer, örneğin hipotetik olarak 100 örneklem seçildiğinde bunlardan en fazla 20 tanesinde tip II hata yapmanın göze alınabildiği anlamına gelir. *İstatistiksel güç* ($1 - \beta$) ise gerçekte

kitlede var olan bir etkinin (H_0 yanlış, H_A doğru) örneklemde de var olduğu çıkarımını yapmak, yani yanlış olan H_0 hipotezini reddetmek ile ilgilidir. İstatistiksel gücün 0.80 olması; hipotetik olarak 100 örneklem seçildiğinde, bunlardan en az 80 tanesinde var olan etkinin tespit edilebilmesi anlamına gelmektedir.

Hipotez Testinin Yönü

Hipotez testleri gerçekleştirilirken, tahmin edici (İng. *Estimator*) ve referans değeri (çoğunlukla yokluk hipotezi değeri) farkının tahmin edicinin standart hatasına bölünmesi ile hesaplanan değer *test istatistiği* olarak adlandırılır (örn. hesaplanan *z* ya da *t* değeri). Test istatistiği, belirli bir dağılım için (örn. standart normal dağılım ya da *t* dağılımı) tip I hata oranına tekabül eden kritik bir değer ile kıyaslanır (örneğin kritik *z* ya da *t* değeri). Hesaplanan değer ve kritik değerin kıyaslanması ile alternatif hipoteze bağlı olarak yokluk hipotezinin reddedilip edilmeyeceğine karar verilir.

Ayrıca, tip I hata oranı belirlenirken hipotez testinin tek ya da çift yönlü olup olmadığı göz önünde bulundurulmalıdır. Tek yönlü hipotez testinde örneklemden elde edilen tahmin edicinin yokluk hipotezinin öne sürdüğü referans değerden daha büyük ya da küçük olduğu öne sürülür. Çift yönlü hipotez testinde ise örneklemden elde edilen tahmin edicinin yokluk hipotezinin öne sürdüğü referans değerden farklı olduğu öne sürülür (küçük olabileceği gibi büyük de olabilir).

Örneğin, tek yönlü hipotez testi için tip I hata 0.05 olarak belirlendiğinde, kritik değerden daha küçük (ya da daha büyük) bir test istatistiği gözlemleme olasılığının 0.05 olduğu anlamına gelir. Kritik değer merkezi dağılımının sadece bir tarafında bulunmaktadır (tek kuyruklu). Ancak, çift yönlü hipotez testi için tip I hata 0.05 olarak belirlendiğinde, test istatistiğinin soldaki kritik değerden daha küçük olma veya sağdaki kritik değerden daha büyük olma olasılığı 0.025 + 0.025 olduğu anlamına gelir. Kritik değer merkezi dağılımın iki tarafında bulunmaktadır (çift kuyruklu).

Yokluk hipotezine yaygın olarak sıfır değeri atanmaktadır (çoğu yazılım programında varsayılan değer). Ancak, yokluk hipotezine pratik anlamda sıfır kabul edilebilecek küçük bir değerin (sınır değeri) atanabildiği durumlar da vardır. Bu mantık doğrultusunda, *Non-inferiority* (pratik anlamda eşdeğer veya daha üstün), *Superority* (pratik anlamda daha üstün) ve *Equivalence* (pratik anlamda eşdeğer) hipotez testleri çoğunlukla tıp ve farmasötik araştırmalarında kullanılmakla birlikte eğitim, davranış ve sosyal bilimlerde de kullanışlı olabilir. Bu tür testlerin ayrıntılarını ve yorumlarını içeren çok sayıda kaynak mevcuttur (örn. Bokai, Hongyue, Xin ve Changyong, 2017; CPMP, 1998, 2001; Serdar, Cihan, Yücel ve Serdar,2021)

Pratik Anlamda En Küçük Anlamlı Etki

Örneklem büyüklüğünün hesaplanması için pratik anlamda anlamlı olabilecek en küçük etkinin ne kadar olacağı belirlenmelidir. En küçük anlamlı etkinin ne olacağına mevcut çalışmalardan, uzmanlardan, raporlardan elde edilen sonuçlara göre karar verilebilir. Örneğin, depresyon hastalarının

belirtilerinde en az ne kadarlık bir iyileşmenin, ya da öğrencilerin başarılarında en az ne kadarlık bir artışın kayda değer bir ilerleme olarak değerlendirilebileceği en küçük anlamlı etki ile ilgilidir.

Önceki çalışmalarda bildirilen etki büyüklüklerinin güç hesaplamalarında kullanılması bazı araştırmacılar tarafından eleştirilmektedir (Bulus ve Koyuncu, 2021; Gelman, 2019). Önceki çalışmalarda bildirilen bir etki pratik anlamda en küçük anlamlı etki olmayabilir. Bununla birlikte, önceki çalışmalarda bildirilen etki büyüklüğü, yeni bir programın en az eski program kadar etkili olup olmadığı veya bir çalışmanın tekrarlanabilirliği araştırılırken kullanılabilir.

Hipotez Testinin Tipi

Hipotez testinin türüne (*t*, *z*, *F*, vb.) bakılmaksızın istatistiksel güç veya örneklem büyüklüğü hesaplama mantığı benzer olsa da, aralarında küçük farklılıklar vardır. Kritik *t* değerini belirlemek (ve istatistiksel gücü hesaplamak) için tip I hataya, serbestlik derecesine ve hipotez testinin yönüne veya türüne ihtiyacımız vardır. Örneklem büyüklüğünü hesaplamak için ise yinelemeli kök bulma algoritmaları kullanılır çünkü kritik *t* değeri serbestlik derecesine, serbestlik derecesi de örneklem büyüklüğüne bağlıdır.

Öte yandan, kritik *z* değerini belirlemek için yalnızca tip I hataya ve hipotez testinin yönüne veya türüne ihtiyacımız vardır. Kritik *z* değeri örneklem büyüklüğünden etkilenmediğinden, örneklem büyüklüğü yinelemeli kök bulma algoritmalarına gerek kalmadan doğrudan formülle hesaplanabilir. Kritik *F* değerini bulmak (ve istatistiksel gücü hesaplamak) için tip I hataya ve serbestlik derecesine (pay ve payda için) ihtiyacımız vardır. Ancak, pay ve payda için serbestlik dereceleri grup veya ölçüm sayısına ve örneklem büyüklüğüne bağlı olduğundan, örneklem büyüklüğünü belirlemek için yinelemeli kök bulma algoritmaları kullanılır.

R ile İstatiksel Güç Analizi

İstatistiksel güç ve örneklem büyüklüğü hesaplamaları için birçok harika program mevcut olsa da (örn., pwr R paketi, Champley ve diğerleri., 2020; G*Power, Erdfelder ve diğerleri., 1996), bir çok platformda erişilebilir çok dilli web uygulamalarının olması pwrss R paketini cazip hale getirmektedir (Bulus, 2023). Paketin R ortamında kurulup aktifleşmesi için aşağıdaki kod grubu kullanılabilir.

```
# Kurulum
install.packages("pwrss")
# Aktifleştirme
library(pwrss)
```

İzleyen bölümlerde önce güç analizi için öncelikle gerekli formüller ve denklemler açıklanacak daha sonra da örnekler üzerinden hesaplamaların nasıl yapılacağı R kodları ile gösterilecektir. Dileyen okuyucular hesaplamaları linkleri verilen web uygulamaları üzerinden gerçekleştirebilirler.

Tek Bir Oranın Bir Sabit ile Karşılaştırılması

Örneklemden elde edilen bir oranın ($\hat{p}$) sabit bir oran (p_0) ile karşılaştırılması z testi ile gerçekleştirilebilir. Gözlem sayısı n olan bir örneklem kullanılarak tahmin edilen $\hat{p}$ değerinin standart hatası $SH(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$ formülü ile ifade edilebilir, ancak, tahmin edici için hesaplanan standart hata tahmin edicinin kendisine bağlı olduğundan istikrarlı olmayan sonuçlar doğurur. Tahmin edilen değer $\hat{p} \cong 0.50$ olduğunda standart hata nispeten büyük, uçlara doğru gidildikçe ise nispeten küçük çıkmaktadır. Bu durumun üstesinden gelmek için Cohen (1988) sinüs fonksiyonunun tersini (İng. *arcsine*) kullanarak oranları dönüştürmüş ve gerekli istatistiksel işlemleri bu dönüştürülen değerler üzerinden yapmayı önermiştir. Tahmin edicinin ters sinüs fonksiyonu dönüşümü

$$\phi_{\hat{p}} = 2\arcsin(\hat{p}) \tag{1}$$

şeklinde yapılır. Bu değerin yaklaşık olarak standart normal dağılımı izlediği ve bundan dolayı standart hatasının $SH(\hat{\phi}) = \sqrt{1/n}$ olarak tanımlandığı düşünülebilir. Test istatistiği ise

$$z = \frac{\phi_{\hat{p}}}{\sqrt{1/n}} \tag{2}$$

şeklinde hesaplanır. Test istatistiği bilindiğinde istatistiksel güç ve örneklem büyüklüğü kolaylıkla hesaplanabilir.

Tek yönlü hipotez testi: Kitledeki p oranının sabit bir p_0 oranından daha küçük ya da daha büyük olduğu düşünüldüğünde, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $p \geq p_0$ (ya da $p \leq p_0$)

H_A: $p < p_0$ (ya da $p > p_0$)

Ters sinüs fonksiyonu dönüşümleri ve test istatistiği Denklem 3-6'daki gibi hesaplanır. Açıklanan denklemler tercihe göre önsel ya da sonsal güç analizini gerçekleştirmek için kullanılabilir. Test istatistiği

$$\phi_{\hat{p}} = 2\arcsin(\hat{p}) \tag{3}$$

$$\phi_{p_0} = 2\arcsin(p_0) \tag{4}$$

$$\hat{h} = \phi_{\hat{p}} - \phi_{p_0} \tag{5}$$

$$z = \frac{\hat{h}}{SH(\hat{h})} = \frac{\phi_{\hat{p}} - \phi_{p_0}}{\sqrt{1/n}} \tag{6}$$

olarak hesaplanır. Denklem 6'da hesaplanan test istatistiği (z) kritik değer (z_k) ile karşılaştırılıp güç analizi yapılabilir. Kritik z_k değeri standart normal dağılımın ters kümülatif yoğunluk fonksiyonunda

(Φ_z^{-1}) tip I hata oranı (α) tanımlanarak Denklem 7'deki gibi elde edilir. İstatiksel güç ($1-\beta$) ise ortalaması z olan standart normal dağılımın kümülatif yoğunluk fonksiyonunda (Φ_z) z ve z_k değerleri tanımlanarak Denklem 8'teki gibi elde edilir.

$$z_k = \Phi_z^{-1}(\alpha; 0) \tag{7}$$

$$1-\beta = 1-\Phi_z(z_k; z) \tag{8}$$

En küçük gerekli örneklem büyüklüğünü hesaplamak için ise, ilk önce, istenen tip I hata (α) ve tip II hata (β) oranlarına tekabül eden tahmini test istatistiği Denklem 9 kullanılarak bulunur ($z = z_\alpha + z_\beta$). Daha sonra, Denklem 6'daki n eşitliğin bir tarafına çekilerek Denklem 10 elde edilir.

$$z_\alpha + z_\beta = \Phi_z^{-1}(\alpha; 0) + \Phi_z^{-1}(\beta; 0) \tag{9}$$

$$n = \frac{\left(z_\alpha + z_\beta\right)^2}{\left(\phi_{\hat{p}} - \phi_{p_0}\right)^2} \tag{10}$$

Örnek: Bir araştırmacı, belli bir ilde ilkokullarda öğrenme güçlüğü çeken çocukların oranının ülkedeki tüm ilkokul öğrencilerini kapsayan kitledeki orandan daha yüksek olup olmadığını merak etmektedir. Bunun için, bir ildeki tüm ilkokul öğrencilerinin bulunduğu listeden rastgele 50 öğrenci seçip öğrenme güçlükleri olup olmadığını test edecektir. Daha sonra öğrenme güçlüğü çeken çocukların örneklemdeki tüm öğrencilere oranını hesaplayacak ve bu oranın kitledeki orandan daha yüksek olup olmadığını bulmaya çalışacaktır. Araştırmacı, tahmin edici $\hat{p}$ değerinin p_0 değerinden düşük olma ihtimalinin olmadığına kanaat getirdiğinden tek yönlü hipotez testi gerçekleştirecektir.

Kitlede öğrenme güçlüğü çeken çocukların oranının 0.06 civarında olduğu bilinmektedir (MEB, 2021). O halde $p_0 = 0.06$ olacaktır. Güç analizine devam etmeden önce araştırmacının tahmin etmesi gereken parametrelerden biri örneklemde ne kadarlık bir oranın beklendiği, bir diğeri ise tip I hatanın ne kadar olması gerektiğidir. Tek bir hipotezin test edildiği çalışmalarda tip I hata oranı yaygınlıkla 0.05 olarak alınır. Beklenen oranın 0.10 olduğunu farz edilirse, 50 katılımcı ile elde edilecek istatistiksel güç `alternative="greater"` olacak şekilde tanımlanarak

```
pwrss.z.prop(p = 0.1, p0 = 0.06, alpha = 0.05,
             n = 50, alternative = "greater")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p = p0
# HA: p > p0
# ------------------------------
#  Statistical power = 0.276
#  n = 50
# ------------------------------
```

```
# Alternative = "greater"
# Non-centrality parameter = 1.051
# Type I error rate = 0.05
# Type II error rate = 0.724
```

şeklinde hesaplanır ve sonuçlar elde edilir. Görülmektedir ki elde edilen güç oranı yaklaşık olarak %27.6'dir. Sosyal bilimlerde istatistiksel güç oranının en az %80 civarında olması yaygın kabul gören bir değerdir. Ancak, ciddi anlamda maddi kaynak, zaman ve personel sınırlılıkları varsa ve katılımcılara ulaşmak problem ise, metodolojik kaliteden ödün vermemek koşuluyla, güç oranının en az %50'nin üzerinde tutulmasına müsamaha gösterilebilir. Bu durumda, çalışmadan elde edilecek sonuçlar tek başına güvenilir olmasa bile, ilerleyen zamanlarda bir meta-analizinde kullanıldığında başka çalışmalarla birlikte anlamlı bir katma değer sağlayabilir.

Sonuç olarak, istatistiksel gücün %80 olması isteniyorsa 50 katılımcıdan daha fazlasına ihtiyaç vardır. Bu şartları sağlayacak örneklem büyüklüğü

```
pwrss.z.prop(p = 0.1, p0 = 0.06, alpha = 0.05,
             power = 0.8, alternative = "greater")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p = p0
# HA: p > p0
# ------------------------------
#  Statistical power = 0.8
#  n = 281
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Örneklem büyüklüğü 281 olarak bulunmuştur. Şayet örneklemden elde edilecek oranın kitledeki orandan daha küçük olduğu düşünülüyorsa `alternative="less"` olacak şekilde tanımlanmalıdır.

Çift yönlü hipotez testi: Kitledeki p oranının sabit p_0 oranına eşit olmadığı düşünülüyorsa, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $p = p_0$

H_A: $p \neq p_0$

Test istatistiği Denklem 6'daki ile aynıdır. Ancak, çift yönlülüğü hesaba katmak için α yerine $\alpha/2$ kullanılır. Bunun da ötesinde, çift yönlü hipotez testi tek yönlü hipotez testlerinden istatistiksel güç

hesaplaması açısından farklılaşmaktadır. Kritik değer (z_k) ve istatistiksel güç ($1-\beta$) Denklem 11 ve 12'de olduğu gibi hesaplanır.

$$z_k = \Phi_z^{-1}(\alpha/2; 0) \tag{11}$$

$$1-\beta = 1-\Phi_z(z_k; z) + \Phi_z(-z_k; z) \tag{12}$$

En küçük gerekli örneklem büyüklüğü, tip I hata oranı daha önce belirtildiği gibi $\alpha/2$ olarak tanımlanarak, tek yönlü hipotez testi kısmında açıklanan Denklem 9 ve 10'da olduğu gibi hesaplanır.

Örnek: Bir önceki örnek bağlamında, öğrenme güçlüğü çeken çocukların oranının tüm kitledeki orana eşit olup olmadığı araştırılsın. Örneklemden elde edilecek tahmini oranının 0.10 olduğunu kabul edilirse, 50 katılımcı ile elde edilecek istatistiksel güç ya da %80 güç için örneklem büyüklüğü `alternative="not equal"` olacak şekilde tanımlanarak

```
pwrss.z.prop(p = 0.1, p0 = 0.06, alpha = 0.05, n = 50,
             alternative = "not equal")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p = p0
# HA: p != p0
# ------------------------------
#  Statistical power = 0.183
#  n = 50
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 1.051
# Type I error rate = 0.05
# Type II error rate = 0.817

pwrss.z.prop(p = 0.1, p0 = 0.06, alpha = 0.05, power = 0.8,
             alternative = "not equal")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p = p0
# HA: p != p0
# ------------------------------
#  Statistical power = 0.8
#  n = 356
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 2.802
```

```
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir.

Şimdiye kadar açıklanan tek yönlü ve çift yönlü hipotez testlerinde $\hat{p} - p_0$ farkı 0'dan büyük/küçük olduğu veya eşit olmadığı sürece yokluk hipotezi reddedilir. Bu fark 0.001 gibi küçük bir değer olsa bile, istatistiksel olarak manidar olduğu sürece, tahmin edilen $\hat{p}$ değerinin referans p_0 değerinden büyük/küçük olduğu veya eşit olmadığı sonucuna ulaşılır. Ancak, çok küçük farklar istatistiksel olarak manidar olsa bile pratikte bir anlam ifade etmeyebilir. Buradan yola çıkarak pratikte anlamlı olmayacak bir fark yani sınır değer tanımlanıp hipotez testleri buna göre gerçekleştirilebilir. Bu sınır değer (İng. *margin*) literatürde genellikle δ sembolü ile gösterilir. Bu mantık çerçevesinde, daha çok tıbbi veya farmasötik araştırmalarda kullanılan, tek yönlü hipotez testinin farklı çeşitleri de bulunmaktadır. İzleyen paragraflarda sınır değerini göz önünde bulunduran *non-inferiority* (pratik anlamda eşdeğerlik ya da aşağı olmama), *superiority* (pratik anlamda üstünlük) ve *equivalence* (pratik anlamda eşdeğerlik) tek yönlü hipotez testleri ele alınmıştır.

Aşağı olmama (*non-inferiority*) ya da üstünlük (*superiority*) hipotez testi: Bir sınır değer göz önünde bulundurularak (δ), kitledeki p oranının sabit p_0 oranından daha aşağı olmama ya da daha üstün olma durumu söz konusu olduğunda kullanılır. Tek yönlü bir hipotez testidir. Aşağı olmama (*non-inferiority*) testi için, yüksek oranının olumlu bir olguyu ifade ettiği durumlarda (örn. üstün zekâlı öğrenci oranı) sınır değeri genellikle negatiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $p - p_0 \leq \delta$

H_A: $p - p_0 > \delta$

ya da yüksek oranının olumsuz bir olguyu ifade ettiği durumlarda (örn. öğrenme güçlüğü olan öğrencilerin oranı) sınır değeri genellikle pozitiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $p - p_0 \geq \delta$

H_A: $p - p_0 < \delta$

Üstünlük (*superiority*) testi için, yüksek oranının olumlu bir olguyu ifade ettiği durumlarda (örn. üstün zekâlı öğrenci oranı) sınır değeri genellikle pozitiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $p - p_0 \leq \delta$

H_A: $p - p_0 > \delta$

ya da yüksek oranının olumsuz bir olguyu ifade ettiği durumlarda (örn. öğrenme güçlüğü olan öğrencilerin oranı) sınır değeri genellikle negatiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $p - p_0 \geq \delta$

H_A: $p - p_0 < \delta$

Hem aşağı olmama (*non-inferiority*) hem de üstünlük (*superiority*) testi için test istatistiği

$$\phi_{p_0+\delta} = 2\arcsin(p_0 + \delta) \tag{13}$$

$$\hat{h} = \phi_{\hat{p}} - \phi_{p_0+\delta} \tag{14}$$

$$z = \frac{\hat{h}}{SH(\hat{h})} = \frac{\phi_{\hat{p}} - \phi_{p_0+\delta}}{\sqrt{1/n}} \tag{15}$$

şeklinde hesaplanır. İstatiksel güç tek yönlü hipotez testinde olduğu gibi Denklem 7 ve 8 kullanılarak hesaplanır. Örneklem büyüklüğü ise Denklem 9'daki eşitlik kullanılarak

$$n = \frac{(z_\alpha + z_\beta)^2}{(\phi_{\hat{p}} - \phi_{p_0+\delta})^2} \tag{16}$$

şeklinde elde edilir.

Örnek: Bir ildeki üstün zekalı çocukların oranının kitledeki orandan ($p_0 = 0.03$) pratik anlamda daha az olup olmadığı araştırılmak istensin. Örneklemden elde edilen değer p_0 değerinden bir miktar daha düşük çıkabilir ($\delta = -0.005$) ancak bunun pratik anlamda bir sorun yaratmayacağı düşünülebilir. Bu durumda, 50 katılımcı ile elde edilecek istatistiksel güç ya da %80 güç için örneklem büyüklüğü `margin=-0.005` ve `alternative="non-inferior"` olacak şekilde tanımlanarak

```
pwrss.z.prop(p = 0.04, p0 = 0.03, margin = -0.005,  alpha = 0.05,
             n = 50, alternative = "non-inferior")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p - p0 <= margin
# HA: p - p0 > margin
# ------------------------------
#  Statistical power = 0.398
#  n = 50
# ------------------------------
# Alternative = "non-inferior"
# Non-centrality parameter = 1.387
# Type I error rate = 0.05
```

```
# Type II error rate = 0.602

pwrss.z.prop(p = 0.04, p0 = 0.03, margin = -0.005, alpha = 0.05,
             power = 0.8, alternative = "non-inferior")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p - p0 <= margin
# HA: p - p0 > margin
# ------------------------------
#  Statistical power = 0.8
#  n = 161
# ------------------------------
# Alternative = "non-inferior"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Buna göre, hesaplama sonucunda 50 katılımcı ile *non-inferiority* hipotez testi için elde edilecek istatistiksel güç %39.8 olarak bulunmuştur. Bu testin %80 istatistiksel güç ile yapılması için en az 161 katılımcıya ihtiyaç vardır.

Şimdi de, bir ildeki üstün zekalı çocukların oranının kitledeki orandan pratik anlamda daha yüksek olup olmadığı araştırılmak istensin. Örneklemden elde edilen değer p_0 değerinden bir miktar daha yüksek çıkabilir ($\delta = 0.005$) ancak bunun pratik anlamda daha yüksek olmadığı düşünülebilir. Diğer taraftan, pratik anlamda fazlalık ya da üstünlük öne sürülmesi için farkın δ'dan daha fazla olması gerekir. Bu durumda, 50 katılımcı ile elde edilecek istatistiksel güç ya da %80 güç için örneklem büyüklüğü `margin=0.005` ve `alternative="superior"` olacak şekilde tanımlanarak

```
pwrss.z.prop(p = 0.04, p0 = 0.03, margin = 0.005,  alpha = 0.05,
             n = 50, alternative = "superior")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p - p0 <= margin
# HA: p - p0 > margin
# ------------------------------
#  Statistical power = 0.012
#  n = 50
# ------------------------------
# Alternative = "superior"
# Non-centrality parameter = -0.615
# Type I error rate = 0.05
```

```
# Type II error rate = 0.988

pwrss.z.prop(p = 0.04, p0 = 0.03, margin = 0.005, alpha = 0.05,
             power = 0.8, alternative = "superior")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p - p0 <= margin
# HA: p - p0 > margin
# ------------------------------
#  Statistical power = 0.8
#  n = 818
# ------------------------------
# Alternative = "superior"
# Non-centrality parameter = -2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Görüldüğü üzere, 50 katılımcı ile *superiority* hipotez testi için elde edilecek istatistiksel güç oldukça düşük çıkmaktadır (%1.2). Ayrıca, bu testin %80 istatistiksel güç ile yapılması için en az 818 katılımcıya ihtiyaç vardır.

Eşdeğerlik (*equivalence*) hipotez testi: Bir sınır değer göz önünde bulundurularak (δ), kitledeki p oranının sabit p_0 oranına eşdeğer olma durumu söz konusu olduğunda kullanılır. İki tane tek yönlü hipotez testi kullanılarak gerçekleştirilir. Bu durumda yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $|p - p_0| \geq \delta$

H_A: $|p - p_0| < \delta$

Test istatistiği

$$\hat{h} = |\phi_{\hat{p}} - \phi_{p_0+\delta}| \tag{17}$$

$$z = \frac{\hat{h}}{SH(\hat{h})} = \frac{|\phi_{\hat{p}} - \phi_{p_0+\delta}|}{\sqrt{1/n}} \tag{18}$$

şeklinde hesaplanır. İstatiksel güç ise iki tane tek yönlü hipotez testi göz önünde bulundurularak

$$z_k = \Phi_z^{-1}(\alpha; 0) \tag{19}$$

$$1 - \beta = 2\big(1 - \Phi_z(z_k; z)\big) - 1 \tag{20}$$

şeklinde hesaplanır. Örneklem büyüklüğü ise

$$z_{\alpha} + z_{\beta/2} = \Phi_z^{-1}(\alpha; 0) + \Phi_z^{-1}(\beta/2; 0) \tag{21}$$

$$n = \frac{\left(z_{\alpha} + z_{\beta/2}\right)^2}{\left(|\phi_{\hat{p}} - \phi_{p_0+\delta}|\right)^2} \tag{22}$$

formülleri kullanılarak hesaplanır.

Örnek: Bir ildeki üstün zekalı çocukların oranının kitle oranına pratik anlamda eşdeğer olup olmadığı araştırılmak istensin. Örneklemden elde edilen oran p_0 değerinden bir miktar düşük ya da yüksek çıkabilir ($\delta = 0.005$) ancak bunun pratik anlamda daha düşük ya da daha yüksek olmadığı düşünülebilir. Bu durumda, 50 katılımcı ile elde edilen istatistiksel güç ya da %80 güç için örneklem büyüklüğü `margin=0.005` ve `alternative="equivalent"` olacak şekilde tanımlanarak

```
pwrss.z.prop(p = 0.04, p0 = 0.03, margin = 0.005, alpha = 0.05,
             n = 50, alternative = "equivalent")
# Approach: Arcsine transformation
# Error: design is not feasible

pwrss.z.prop(p = 0.04, p0 = 0.03, margin = 0.005, alpha = 0.05,
             power = 0.80, alternative = "equivalent")
# Approach: Arcsine transformation
# One proportion compared to a constant (one sample z test)
# H0: |p - p0| >= margin
# HA: |p - p0| < margin
# ------------------------------
#  Statistical power = 0.8
#  n = 1132
# ------------------------------
# Alternative = "equivalent"
# Non-centrality parameter = -2.926
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Buna göre, 50 katılımcı ile güç oranını hesaplamak mümkün değildir, ancak, %80 güç ile hipotez testi gerçekleştirilmek istenirse en az 1132 kişi gerekmektedir.

İki Oran Farkı

Bir önceki bölümde tek bir örneklemden elde edilen oranın sabit bir orana karşı test edilmesi süreci ele alındı. Standart hataya sadece tek bir oran katkıda bulunmaktaydı çünkü karşılaştırılan referans değer bir sabit olup örnekleme bağlı olarak değişmemekteydi. Diğer taraftan, bir örneklemdeki iki farklı gruba ya da iki farklı örnekleme ait iki oran da karşılaştırılabilir. Bu durumda karşılaştırılan oranların ikisi de tahmin edilmiş oranlar olduğu için standart hataya ikisinin de örneklem hatası katkıda

bulunur. Tahmin edilen değerlerin varyansları sırasıyla $\hat{p}_1(1-\hat{p}_1)$ ve $\hat{p}_2(1-\hat{p}_2)$ olur. O halde $\hat{p}_1 - \hat{p}_2$ farkının standart hatası

$$SH(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{23}$$

formülü ile hesaplanabilir. Ancak, bir önceki bölümde bahsedildiği gibi standart hata tahmin edicilerin ($\hat{p}_1$ ve $\hat{p}_2$) kendilerine bağlıdır. Bu sorunun üstesinden gelmek için ters sinüs fonksiyonu dönüşümleri ve standart hata

$$\phi_{\hat{p}_1} = 2\arcsin(\hat{p}_1) \tag{24}$$

$$\phi_{\hat{p}_2} = 2\arcsin(\hat{p}_2) \tag{25}$$

$$\hat{h} = \phi_{\hat{p}_1} - \phi_{\hat{p}_2} \tag{26}$$

$$SH(\hat{h}) = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{27}$$

şeklinde ifade edilir.

Tek yönlü hipotez testi: Kitledeki bir gruba ait p_1 oranının başka bir gruba ait p_2 oranından daha küçük ya da daha büyük olduğu düşünülüyorsa, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $p_1 \geq p_2$ (ya da $p_1 \leq p_2$)

H_A: $p_1 < p_2$ (ya da $p_1 > p_2$)

Test istatistiği

$$z = \frac{\hat{h}}{SH(\hat{h})} = \frac{\phi_{\hat{p}_1} - \phi_{\hat{p}_2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{28}$$

şeklinde hesaplanır. İstatiksel gücü hesaplamak için Denklem 7 ve 8 kullanılır. Denklem 9'daki işlem yapıldıktan sonra, örneklem büyüklüğü ise

$$n_2 = \frac{(z_\alpha + z_\beta)^2}{(\phi_{\hat{p}_1} - \phi_{\hat{p}_2})^2}\left(\frac{\kappa+1}{\kappa}\right) \tag{29}$$

şeklinde hesaplanır. Denklem 29'da $\kappa = n_1/n_2$ şeklinde tanımlanmıştır. Burada n_1 birinci grubun, n_2 ise ikinci grubun örneklem büyüklüğüdür. O halde $n_1 = n_2\kappa$ şeklinde hesaplanabilir.

Örnek: Belli bir ildeki ilkokullarda öğrenme güçlüğü yaşayan erkek çocukların oranının öğrenme güçlüğü yaşayan kız çocuklarının oranından daha yüksek olup olmadığını bulmaya çalışan bir

araştırmacı her bir gruptan 50'şer kişi olmak üzere toplam 100 kişiden veri toplamayı amaçlamaktadır. Öğrenme güçlüğü çeken erkek çocuklarda beklenen oran $p_1 = 0.08$, kız çocuklarında ise $p_2 = 0.06$ olduğu farz edilsin. İstatistiksel güç ve örneklem büyüklüğü

```
pwrss.z.2props(p1 = 0.08, p2 = 0.06, alpha = 0.05,
               kappa = 1, n2 = 50, alternative = "greater")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: p1 = p2
# HA: p1 > p2
# ------------------------------
#  Statistical power = 0.105
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 0.393
# Type I error rate = 0.05
# Type II error rate = 0.895

pwrss.z.2props(p1 = 0.08, p2 = 0.06, alpha = 0.05,
               kappa = 1, power = 0.8, alternative = "greater")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: p1 = p2
# HA: p1 > p2
# ------------------------------
#  Statistical power = 0.8
#  n1 = 2003
#  n2 = 2003
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. `kappa=1` argümanı erkek katılımcıların kız katılımcılara oranını belirtmektedir (`n1/n2`). Yukarıdaki çıktıdan görülmektedir ki sadece 100 katılımcıdan veri toplamak yeterli olmamaktadır çünkü istatistiksel güç %10.5 civarındadır. %80 güç ile bu test gerçekleştirilmek istenirse her bir gruptan 2003 katılımcıdan veri toplanması gerekmektedir.

Çift yönlü hipotez testi: Kitledeki bir grubun p_1 oranının başka bir grubun p_2 oranından farklı olduğu düşünülüyorsa, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $p_1 = p_2$

H_A: $p_1 \neq p_2$

Test istatistiği denklem 28'de olduğu gibi hesaplanır. İstatiksel gücü hesaplamak için Denklem 11 ve 12 kullanılır. Denklem 9'da α yerine $\alpha/2$ yazılıp gerekli işlem yapıldıktan sonra, ikinci grup için en küçük gerekli örneklem büyüklüğü

$$n_2 = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2}{\left(\phi_{\hat{p}_1} - \phi_{\hat{p}_2}\right)^2}\left(\frac{\kappa + 1}{\kappa}\right) \tag{30}$$

şeklinde hesaplanır. Benzer şekilde birinci grubun örneklem büyüklüğü $n_2 = n_1\kappa$ kullanılarak hesaplanabilir.

Örnek: Bir önceki örnek bağlamında, öğrenme güçlüğü yaşayan erkek çocukların oranının öğrenme güçlüğü yaşayan kız çocukların oranına eşit olmadığını bulmaya çalışan bir araştırmacı çift yönlü hipotez testini gerçekleştirecektir. İstatistiksel güç ve örneklem büyüklüğü `alternative="not equal"` olacak şekilde tanımlanarak

```
pwrss.z.2props(p1 = 0.08, p2 = 0.06, alpha = 0.05,
               kappa = 1, n2 = 50, alternative = "not equal")
# Approach: Arcsine transformation
# Difference between two proportions (independent samples z test)
# H0: p1 = p2
# HA: p1 != p2
# ------------------------------
#  Statistical power = 0.068
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 0.393
# Type I error rate = 0.05
# Type II error rate = 0.932

pwrss.z.2props(p1 = 0.08, p2 = 0.06, alpha = 0.05,
               kappa = 1, power = 0.8, alternative = "not equal")
# Approach: Arcsine transformation
# Difference between two proportions (independent samples z test)
# H0: p1 = p2
# HA: p1 != p2
# ------------------------------
```

```
#  Statistical power = 0.8
#  n1 = 2543
#  n2 = 2543
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 2.802
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Buna göre, 50 katılımcı ile hipotez testinin istatistiksel gücü %6.8 çıkmaktadır. Bu testin %80 istatistiksel güç ile yapılması için her bir grupta en az 2543 katılımcıya ihtiyaç vardır. Daha önce bahsedilen *non-inferiority*, *superiority* ve *equivalence* hipotez testleri burada da kurulabilir.

Aşağı olmama (*non-inferiority*) ya da üstünlük (*superiority*) hipotez testi: Bir sınır değer göz önünde bulundurularak (δ), kitledeki bir gruba ait p_1 oranının diğer bir gruptaki p_2 oranından daha aşağı olmama ya da daha üstün olma durumu söz konusu ise kullanılır. Aşağı olmama (*non-inferiority*) testi için, yüksek oranın olumlu bir olguyu ifade ettiği durumlarda (örn. üstün zekâlı öğrenci oranı) sınır değeri genellikle negatiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $p_1 - p_2 \leq \delta$

H_A: $p_1 - p_2 > \delta$

ya da yüksek oranının olumsuz bir olguyu ifade ettiği durumlarda (örn. öğrenme güçlüğü olan öğrencilerin oranı) sınır değeri genellikle pozitiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $p_1 - p_2 \geq \delta$

H_A: $p_1 - p_2 < \delta$

Üstünlük (*superiority*) testi için, yüksek oranının olumlu bir olguyu ifade ettiği durumlarda (örn. üstün zekâlı öğrenci oranı) sınır değeri genellikle pozitiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $p_1 - p_2 \leq \delta$

H_A: $p_1 - p_2 > \delta$

ya da yüksek oranının olumsuz bir olguyu ifade ettiği durumlarda (örn. öğrenme güçlüğü olan öğrencilerin oranı) sınır değeri genellikle negatiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $p_1 - p_2 \geq \delta$

H_A: $p_1 - p_2 < \delta$

Hem aşağı olmama (*non-inferiority*) hem de üstünlük (*superiority*) testi için hesaplanan test istatistiği

$$z = \frac{\hat{h}}{SH(\hat{h})} = \frac{\phi_{\hat{p}_1} - \phi_{\hat{p}_2+\delta}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{31}$$

şeklinde ifade edilir. İstatiksel gücü hesaplamak için Denklem 7 ve 8 kullanılır. Denklem 9'daki işlem yapıldıktan sonra, ikinci grup için en küçük gerekli örneklem büyüklüğü

$$n_2 = \frac{(z_\alpha + z_\beta)^2}{(\phi_{\hat{p}_1} - \phi_{\hat{p}_2+\delta})^2}\left(\frac{\kappa + 1}{\kappa}\right) \tag{32}$$

şeklinde hesaplanır. Benzer şekilde birinci grup için örneklem büyüklüğü $n_1 = n_2\kappa$ kullanılarak hesaplanabilir.

Örnek: Bir araştırmacı bir ildeki üstün zekâlı erkek çocukların oranının üstün zekâlı kız çocukların oranından pratik anlamda daha düşük olup olmadığını araştırmak istemektedir. Birinci grubun oranının ikinci grubun oranından farkı δ=-0.005 olsa bile pratik anlamda p_1'in p_2'den daha az olmadığı farz edilsin. Bu durumda, her bir gruptan 50 katılımcı olacak şekilde istatistiksel güç ya da %80 istatistiksel güç için örneklem büyüklüğü `margin=-0.005` ve `alternative="non-inferior"` olacak şekilde tanımlanarak

```
pwrss.z.2props(p1 = 0.04, p2 = 0.02, alpha = 0.05, margin = -0.005,
               kappa = 1, n2 = 50, alternative = "non-inferior")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: p1 - p2 <= margin
# HA: p1 - p2 > margin
# ------------------------------
#  Statistical power = 0.366
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "non-inferior"
# Non-centrality parameter = 1.302
# Type I error rate = 0.05
# Type II error rate = 0.634
```

```
pwrss.z.2props(p1 = 0.04, p2 = 0.02, alpha = 0.05,  margin = -0.005,
               kappa = 1, power = 0.8, alternative = "non-inferior")
# Approach: Arcsine transformation
# Difference between two proportions
```

```
# (independent samples z test)
# H0: p1 - p2 <= margin
# HA: p1 - p2 > margin
# ------------------------------
#  Statistical power = 0.8
#  n1 = 183
#  n2 = 183
# ------------------------------
# Alternative = "non-inferior"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Buna göre, her bir grupta 50 katılımcı ile *non-inferiority* hipotez testi için elde edilecek istatistiksel güç %36.6 çıkmaktadır. Bu testin %80 istatistiksel güç ile yapılması için her bir gruptan en az 183 katılımcıya ihtiyaç olduğu görülmektedir.

Şimdi de, bir ildeki üstün zekâlı erkek çocukların oranının üstün zekâlı kız çocukların oranından pratik anlamda daha yüksek olup olmadığı araştırılmak istensin. İki oran arasındaki fark δ=0.01'ten daha büyük olduğunda pratik anlamda p_1'in p_2'den daha fazla olduğunu farz edilsin. Yine her bir gruptan 50 katılımcı olacak şekilde istatistiksel güç ya da %80 güç için örneklem büyüklüğü `margin=0.01` ve `alternative="superior"` olacak şekilde tanımlanarak

```
pwrss.z.2props(p1 = 0.04, p2 = 0.02, alpha = 0.05, margin = 0.01,
               kappa = 1, n2 = 50, alternative = "superior")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: p1 - p2 <= margin
# HA: p1 - p2 > margin
# ------------------------------
#  Statistical power = 0.02
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "superior"
# Non-centrality parameter = -0.113
# Type I error rate = 0.05
# Type II error rate = 0.961

pwrss.z.2props(p1 = 0.04, p2 = 0.02, alpha = 0.05,  margin = 0.01,
               kappa = 1, power = 0.8, alternative = "superior")
# Approach: Arcsine transformation
# Difference between two proportions
```

```
# (independent samples z test)
# H0: p1 - p2 <= margin
# HA: p1 - p2 > margin
# ------------------------------
#  Statistical power = 0.8
#  n1 = 1866
#  n2 = 1866
# ------------------------------
# Alternative = "superior"
# Non-centrality parameter = -2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Buna göre, her bir grupta elli katılımcı ile *superiority* hipotez testinin gücü %2 çıkmaktadır. Bu testin %80 istatistiksel güç ile yapılması için her bir grupta en az 1866 katılımcıya ihtiyaç vardır. Dikkat edilecek olunursa *non-inferiority* ve *superiority* arasındaki tek fark δ'nın nasıl tanımlandığıdır.

Eşdeğerlik (*equivalence*) hipotez testi: Bir sınır değer göz önünde bulundurularak (δ), kitledeki bir gruba ait p_1 oranının diğer bir gruptaki p_2 oranına eşdeğer ya da denk olma durumu araştırılıyorsa, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $|p_1 - p_2| \geq \delta$

H_A: $|p_1 - p_2| < \delta$

Test istatistiği

$$z = \frac{\hat{h}}{SH(\hat{h})} = \frac{|\phi_{\hat{p}_1} - \phi_{\hat{p}_2+\delta}|}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{33}$$

şeklinde hesaplanır. İstatiksel gücü hesaplamak için Denklem 19 ve 20 kullanılır. Denklem 9'daki işlem yapıldıktan sonra, örneklem büyüklüğü ise

$$n_2 = \frac{\left(z_\alpha + z_\beta\right)^2}{\left(|\phi_{\hat{p}_1} - \phi_{\hat{p}_2+\delta}|\right)^2}\left(\frac{\kappa + 1}{\kappa}\right) \tag{34}$$

şeklinde hesaplanır. Benzer şekilde, $n_1 = n_2\kappa$ kullanılarak hesaplanabilir.

Örnek: Bir ildeki üstün zekâlı erkek çocukların oranının üstün zekâlı kız çocukların oranına pratik anlamda eşdeğer olup olmadığı araştırılsın. İki oran arasındaki fark 0.01'den daha küçük ya da -0.01'ten daha büyük ($\delta = 0.01$) olduğunda pratik anlamda p_1'in p_2'e eşdeğer olduğunu farz edilsin. Yine her bir gruptan 50 katılımcı olacak şekilde istatistiksel güç ya da %80 güç için örneklem büyüklüğü `margin=0.01` ve `alternative="equivalent"` olacak şekilde tanımlanarak

```
pwrss.z.2props(p1 = 0.04, p2 = 0.02, alpha = 0.05, margin = 0.01,
               n2 = 50, kappa = 1, alternative = "equivalent")
# Approach: Arcsine transformation
# Error: design is not feasible

pwrss.z.2props(p1 = 0.04, p2 = 0.02, alpha = 0.05,  margin = 0.01,
               kappa = 1, power = 0.8, alternative = "equivalent")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: |p1 - p2| >= margin
# HA: |p1 - p2| < margin
# ------------------------------
#  Statistical power = 0.8
#  n1 = 2585
#  n2 = 2585
# ------------------------------
# Alternative = "equivalent"
# Non-centrality parameter = -2.926
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Burada, istatistiksel gücü hesaplamak için her bir grupta 50 kişinin yeterli olmayacağı verilen hatadan görülmektedir (`Error: design is not feasible`). Ayrıca, %80 istatistiksel güç ile eşdeğerlik testi için her bir gruptan 2585 katılımcı gerekmektedir.

Tek Ortalamanın Bir Sabitle Karşılaştırması

Kitlede ortalaması μ ve varyansı σ^2 olan rassal bir X değişkeni olduğu farz edilsin. Büyüklüğü n olan bir örneklemden X değişkeni için $x_1, x_2, x_3, \ldots, x_n$ değerleri gözlensin. X'in ortalamasının tahmin edicisi

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{35}$$

ve X'in varyansının tahmin edicisi

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \tag{36}$$

şeklinde hesaplanır. $\hat{\mu}$'in standart hatası ise

$$SH(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{n}} \tag{37}$$

şeklinde ifade edilir.

Tek yönlü hipotez testi: Kitledeki μ ortalamasının sabit bir değerden (μ_0) daha küçük ya da daha büyük olduğu düşünülüyorsa, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi kurulur:

H_0: $\mu \geq \mu_0$ (ya da $\mu \leq \mu_0$)

H_A: $\mu < \mu_0$ (ya da $\mu > \mu_0$)

Test istatistiği

$$z = \frac{\hat{\mu} - \mu_0}{SH(\hat{\mu})} = \frac{\hat{\mu} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}} \tag{38}$$

şeklinde hesaplanır. İstatistiksel güç hesaplaması için Denklem 7 ve 8 kullanılır. Daha sonra Denklem 9'daki işlem uygulandıktan sonra, en küçük gerekli örneklem büyüklüğü

$$n = \frac{(z_\alpha + z_\beta)^2 \hat{\sigma}^2}{(\hat{\mu} - \mu_0)^2} \tag{39}$$

şeklinde hesaplanır.

Örnek: Bir araştırmacı, COVID-19 döneminde üniversite öğrencilerinin depresyon seviyesinin ortalama düzey kabul edilebilecek 21 puandan daha yüksek olup olmadığını bulmak istemektedir. Araştırma için kullanılacak Beck Depresyon Envanteri (BDE; Beck, Ward, Mendelson, Mock ve Erbaugh, 1961) 21 maddeden oluşmaktadır. Hisli (1989) 21 puan alan üniversite öğrencilerinin orta düzey seviyesinde depresyon belirtileri gösterdiklerini ifade etmektedir. O halde μ_0 = 21 olarak tanımlanabilir. Ayrıca, Hisli (1989) üniversite öğrencilerinin depresyon puanlarının standart sapmasını 6.75 olarak raporlamıştır. Ortalama düzeyden iki birimlik artışın (21 + 2), en küçük anlamlı fark olduğu farz edilirse, 50 katılımcı ile elde edilen istatistiksel güç ya da %80 güç için gerekli örneklem büyüklüğü `alternative="greater"` olacak şekilde tanımlanarak

```
pwrss.z.mean(mu = 23, mu0 = 21, sd = 6.75,
             alpha = 0.05, n = 50, alternative = "greater")
# One mean compared to a constant
# (one sample z test)
# H0: mu = mu0
# HA: mu > mu0
# ------------------------------
#  Statistical power = 0.674
#  n = 50
```

```
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 2.095
# Type I error rate = 0.05
# Type II error rate = 0.326

pwrss.z.mean(mu = 23, mu0 = 21, sd = 6.75,
             alpha = 0.05, power = 0.8, alternative = "greater")
# One mean compared to a constant
# (one sample z test)
# H0: mu = mu0
# HA: mu > mu0
# ------------------------------
#  Statistical power = 0.8
#  n = 71
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Görüldüğü üzere, 50 katılımcı ile tek yönlü hipotez testinin gücü %67.4 olarak bulunmuştur. Ayrıca daha fazla katılımcıdan veri toplamak mümkün ise, hipotez testini %80 istatistiksel güç ile gerçekleştirmek için en az 71 katılımcıya ihtiyaç olduğu görülmektedir.

Çift yönlü hipotez testi: Kitledeki μ ortalamasının sabit bir değere (μ_0) eşit olmadığı araştırıldığında, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi kurulur:

H_0: $\mu = \mu_0$

H_A: $\mu \neq \mu_0$

Test istatistiği Denklem 38'deki gibi hesaplanır. İstatistiksel güç hesaplaması için ise Denklem 7 ve 8 kullanılır ve örneklem büyüklüğü

$$n = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2 \hat{\sigma}^2}{(\hat{\mu} - \mu_0)^2} \tag{40}$$

şeklinde hesaplanır.

Örnek: Bir araştırmacı, COVID-19 döneminde üniversite öğrencilerinin depresyon seviyesinin ortalama düzey kabul edilebilecek 21 puandan farklı olduğunu kanıtlanmaya çalışmaktadır. Bir önceki örnekte olduğu gibi, en küçük anlamlı farkın 2 birim olduğu farz edilirse (yani depresyon düzeyi 19 birim olabileceği gibi 23 birim de olabilir) 50 katılımcı ile elde edilecek istatistiksel güç ya da %80 güç için örneklem büyüklüğü `alternative="not equal"` olacak şekilde tanımlanarak

```
pwrss.z.mean(mu = 23, mu0 = 21, sd = 6.75,
             alpha = 0.05, n = 50, alternative = "not equal")
# One mean compared to a constant
# (one sample z test)
# H0: mu = mu0
# HA: mu != mu0
# ------------------------------
#  Statistical power = 0.554
#  n = 50
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 2.095
# Type I error rate = 0.05
# Type II error rate = 0.446

pwrss.z.mean(mu = 23, mu0 = 21, sd = 6.75,
             alpha = 0.05, power = 0.8, alternative = "not equal")
# One mean compared to a constant
# (one sample z test)
# H0: mu = mu0
# HA: mu != mu0
# ------------------------------
#  Statistical power = 0.8
#  n = 90
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 2.802
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Görüldüğü üzere, 50 katılımcı ile çift yönlü hipotez testinin gücü %55.4 olarak bulunmuştur. Ayrıca daha fazla katılımcıdan veri toplamak mümkün ise, hipotez testini %80 istatistiksel güç ile gerçekleştirmek için en az 90 katılımcıya ihtiyaç vardır.

Aşağı olmama (*non-inferiority*) ya da üstünlük (*superiority*) hipotez testi: Bir sınır değer (δ) göz önünde bulundurularak, kitledeki μ ortalamasının sabit bir değerinden (μ_0) daha aşağı olmama ya da daha üstün olma durumu söz konusu olduğunda kullanılır. Aşağı olmama (*non-inferiority*) testi için, değişkenin yüksek değerlerinin olumlu bir olguyu ifade ettiği durumlarda (örn. başarı testi puanı) sınır değeri genellikle negatiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\mu - \mu_0 \leq \delta$

H_A: $\mu - \mu_0 > \delta$

ya da değişkenin yüksek değerlerinin olumsuz bir olguyu ifade ettiği durumlarda sınır değeri genellikle pozitiftir (örn. depresyon puanı) ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\mu - \mu_0 \geq \delta$

H_A: $\mu - \mu_0 < \delta$

Üstünlük (*superiority*) testi için, değişkenin yüksek değerlerinin olumlu bir olguyu ifade ettiği durumlarda (örn. başarı testi puanı) sınır değeri genellikle pozitiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\mu - \mu_0 \leq \delta$

H_A: $\mu - \mu_0 > \delta$

ya da değişkenin yüksek değerlerinin olumsuz bir olguyu ifade ettiği durumlarda (örn. depresyon puanı) sınır değeri genellikle negatiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\mu - \mu_0 \geq \delta$

H_A: $\mu - \mu_0 < \delta$

Hem aşağı olmama (*non-inferiority*) hem de üstünlük (*superiority*) testleri için hesaplanan test istatistiği

$$z = \frac{\hat{\mu} - \mu_0 - \delta}{SH(\hat{\mu})} = \frac{\hat{\mu} - \mu_0 - \delta}{\frac{\hat{\sigma}}{\sqrt{n}}} \tag{41}$$

şeklinde ifade edilir. İstatistiksel güç hesaplaması için Denklem 7 ve 8 kullanılır. En küçük gerekli örneklem büyüklüğü

$$n = \frac{\left(z_\alpha + z_\beta\right)^2 \hat{\sigma}^2}{(\hat{\mu} - \mu_0 - \delta)^2} \tag{42}$$

şeklinde hesaplanır.

Örnek: Bir araştırmacı, COVID-19 döneminde ortaokul öğrencilerinin psikolojik sağlamlık seviyelerinin COVID-19 öncesinde yayınlanan bir makaledeki değerden pratik anlamda daha az olup olmadığını bulmaya çalışmaktadır. Alternatif hipotezin öne sürdüğü ortalama ile yokluk hipotezi değeri farkının pratik anlamda sınır değeri δ = -2 olarak belirlenmiştir. Daha önce yapılan çalışmada psikolojik sağlamlığın ortalama değeri 49 birim ve psikolojik sağlamlık puanlarının standart sapması 7.59 olarak rapor edilmiştir (Arslan, 2015). Örneklemden elde edilecek beklenen ortalamanın 51 birim olduğu farz edilirse, 50 katılımcı ile elde edilecek istatistiksel güç ya da %80 güç için örneklem büyüklüğü `margin=-2` ve `alternative="non-inferior"` olacak şekilde tanımlanarak

```
pwrss.z.mean(mu = 51, mu0 = 49, sd = 7.59, margin = -2,
             alpha = 0.05, n = 50, alternative = "non-inferior")
# One mean compared to a constant
# (one sample z test)
# H0: mu - mu0 <= margin
# HA: mu - mu0 > margin
# ------------------------------
#  Statistical power = 0.981
#  n = 50
# ------------------------------
# Alternative = "non-inferior"
# Non-centrality parameter = 3.727
# Type I error rate = 0.05
# Type II error rate = 0.019

pwrss.z.mean(mu = 51, mu0 = 49, sd = 7.59, margin = -2,
             alpha = 0.05, power = 0.8, alternative = "non-inferior")
# One mean compared to a constant
# (one sample z test)
# H0: mu - mu0 <= margin
# HA: mu - mu0 > margin
# ------------------------------
#  Statistical power = 0.8
#  n = 23
# ------------------------------
# Alternative = "non-inferior"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır. Sonuç olarak, 50 katılımcı ile *non-inferiority* hipotez testinin gücü %98.1 olarak bulunmuştur. Ayrıca, hipotez testini %80 istatistiksel güç ile gerçekleştirmek için ise sadece 23 katılımcı yeterlidir.

Eşdeğerlik (*equivalence*) hipotez testi: Bir sınır değer göz önünde bulundurularak (δ), kitledeki μ ortalamasının sabit bir değere (μ_0) eşdeğer olması söz konusu ise, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $|\mu - \mu_0| \geq \delta$

H_A: $|\mu - \mu_0| < \delta$

Test istatistiği

$$z = \frac{|\hat{\mu} - \mu_0| - \delta}{SH(\hat{\mu})} = \frac{|\hat{\mu} - \mu_0| - \delta}{\frac{\hat{\sigma}}{\sqrt{n}}} \tag{43}$$

şeklinde hesaplanır. İstatistiksel güç hesaplaması için Denklem 17 ve 18 kullanılır ve örneklem büyüklüğü ise

$$n = \frac{\left(z_\alpha + z_{\beta/2}\right)^2 \hat{\sigma}^2}{(|\hat{\mu} - \mu_0| - \delta)^2} \tag{44}$$

şeklinde hesaplanır.

Örnek: Bir araştırmacı, ortaokul öğrencilerinin psikolojik sağlamlık seviyelerinin COVID-19 öncesinde yayınlanan bir makaledeki değere eşdeğer olup olmadığı bulmaya çalışmaktadır. Burada δ yine pratik anlamda farkın sınır değerdir. Eşdeğerlik hipotez testinde δ = 1 olması şu anlama gelir; örneklemden elde edilen $|\hat{\mu} - \mu_0|$ farkı 1 birimden az olmalıdır ki eşdeğerlik öne sürülebilsin. O halde, 50 katılımcı ile elde edilen istatistiksel güç ya da %80 güç için örneklem büyüklüğü `margin=1` ve `alternative="equivalent"` olacak şekilde tanımlanarak

```
pwrss.z.mean(mu = 51, mu0 = 49, sd = 7.59, margin = 1,
             alpha = 0.05, n = 50, alternative = "equivalent")
# Error: design is not feasible

pwrss.z.mean(mu = 51, mu0 = 49, sd = 7.59, margin = 1,
             alpha = 0.05, power = 0.8, alternative = "equivalent")
# One mean compared to a constant
# (one sample z test)
# H0: |mu - mu0| >= margin
# HA: |mu - mu0| < margin
# ------------------------------
#  Statistical power = 0.8
#  n = 494
# ------------------------------
# Alternative = "equivalent"
# Non-centrality parameter = 2.926
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Yukarıdaki çıktıda verilen hata (`Error: design is not feasible`) istatistiksel gücü 50 kişi ile belirleyebilmenin mümkün olmadığını göstermektedir. Öte yandan, %80 güç ile eşdeğerlik hipotez testini gerçekleştirmek için en az 494 kişi gerektiği de hesaplanmıştır.

İki Ortalama Farkı

Bir zaman kesitinde iki gruba ait ölçümlerin ortalamaları veya aynı gruba ait iki farklı zamanda yapılan ölçümlerin ortalamaları karşılaştırılmak istenebilir. Bir zaman kesitinde iki gruba ait ortalamaları karşılaştırmak için bağımsız örneklemler t testi, aynı grubun iki farklı zamana ait ortalamaları karşılaştırmak için ise bağımlı örneklemler t testi kullanılır.

Bağımsız örneklemler t testi: Birinci grup için ortalaması μ_1 ve varyansı σ_1^2 olan rassal bir X_1 değişkeni, ikinci grup için ise ortalaması μ_2 ve varyansı σ_2^2 olan rassal bir X_2 değişkeni düşünülsün. Büyüklüğü n_1 olan bir örneklemde X_1 ile ilgili $x_{11}, x_{12}, x_{13}, \dots, x_{1n_1}$ değerleri, büyüklüğü n_2 olan bir örneklemde X_2 ile ilgili $x_{21}, x_{22}, x_{23}, \dots, x_{2n_2}$ değerleri gözlensin. X_1 ve X_2'in ortalamalarının tahmin edicileri

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i} \tag{45}$$

$$\hat{\mu}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i} \tag{46}$$

varyanslarının tahmin edicileri ise

$$\hat{\sigma}_1{}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} (x_{1i} - \hat{\mu}_1)^2 \tag{47}$$

$$\hat{\sigma}_2{}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2} (x_{2i} - \hat{\mu}_2)^2 \tag{48}$$

şeklinde hesaplanır. $\hat{\mu}_1 - \hat{\mu}_2$'in standart hatası ise

$$SH(\hat{\mu}_1 - \hat{\mu}_2) = \sqrt{\frac{\hat{\sigma}_1{}^2}{n_1} + \frac{\hat{\sigma}_2{}^2}{n_2}} \tag{49}$$

formülünden elde edilir.

Tek yönlü hipotez testi: Kitledeki bir grubun ortalamasının (μ_1) diğer grubun (μ_2) ortalamasından daha küçük ya da daha büyük olduğu düşünülüyorsa, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi kurulur:

H_0: $\mu_1 \geq \mu_2$ (ya da $\mu_1 \leq \mu_2$)

H_A: $\mu_1 < \mu_2$ (ya da $\mu_1 > \mu_2$)

Test istatistiği

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{SH(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^{\,2}}{n_1} + \frac{\hat{\sigma}_2^{\,2}}{n_2}}} \tag{50}$$

şeklinde hesaplanır. İstatiksel gücü hesaplamak için

$$v = n_1 + n_2 - 2 \tag{51}$$

$$t_k = \Phi_t^{-1}(\alpha, v; 0) \tag{52}$$

$$1 - \beta = 1 - \Phi_t(t_k, v; t) \tag{53}$$

denklemleri kullanılır. Burada v serbestlik derecesini ifade etmektedir ve Denklem 51'de olduğu gibi hesaplanır. İkinci grup için örneklem büyüklüğünü hesaplamak için

$$t_\alpha + t_\beta = \Phi_t^{-1}(\alpha, v; 0) + \Phi_t^{-1}(\beta, v; 0) \tag{54}$$

$$n_2 = (t_\alpha + t_\beta)^2 \left(\frac{\frac{\hat{\sigma}_1^{\,2}}{\kappa} + \hat{\sigma}_2^{\,2}}{(\hat{\mu}_1 - \hat{\mu}_2)^2} \right) \tag{55}$$

denklemleri kullanılır. Birinci grubun örneklem büyüklüğü, $\kappa = n_1/n_2$ şeklinde tanımlandığından, $n_1 = n_2\kappa$ kullanılarak hesaplanabilir. Standardize ortalama farkı olarak Cohen d kullanılarak istatistiksel güç ve örneklem büyüklüğü hesaplamaları yapılacaksa $\hat{\mu}_1 = d$, $\hat{\mu}_2 = 0$, $\hat{\sigma}_1^{\,2} = 1$ ve $\hat{\sigma}_2^{\,2} = 1$ şeklindeki tanımlanmalıdır (varsayılan ayarlar bu şekildedir).

Örnek: Bir araştırmacı, COVID-19 döneminde üniversitelerde okuyan kız öğrencilerin depresyon seviyelerinin erkeklere kıyasla daha yüksek olduğunu öne sürmektedir. Daha önceki çalışmalardan depresyon puanlarının standart sapmasının 6.75 olduğunu raporlanmıştır (Hisli, 1989). Bu değer, iki grubun verisi birlikte kullanılarak hesaplandığından havuzlamış (birleştirilmiş) standart sapma olarak düşünülebilir. Yine en küçük anlamlı ortalama farkının 2 birim olduğu farz edilirse (kızlar için 26 erkekler için ise 24), 50'şer katılımcı ile elde edilecek istatistiksel güç ya da %80 güç için örneklem büyüklüğü `alternative="greater"` olacak şekilde tanımlanarak

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, kappa = 1,
               alpha = 0.05, n = 50, alternative = "greater")
# Difference between two means
# (independent samples t test)
# H0: mu1 = mu2
# HA: mu1 > mu2
# ------------------------------
```

```
#  Statistical power = 0.431
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "greater"
# Degrees of freedom = 98
# Non-centrality parameter = 1.481
# Type I error rate = 0.05
# Type II error rate = 0.569

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, kappa = 1,
               alpha = 0.05, power = 0.8, alternative = "greater")
# Difference between two means
# (independent samples t test)
# H0: mu1 = mu2
# HA: mu1 > mu2
# ------------------------------
#  Statistical power = 0.8
#  n1 = 142
#  n2 = 142
# ------------------------------
# Alternative = "greater"
# Degrees of freedom = 281.2
# Non-centrality parameter = 2.493
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır. `kappa=1` argümanı kız katılımcıların erkek katılımcılara oranını belirtmektedir (`n1/n2`). `sd1` argümanı birinci grubun standart sapmasını temsil ettiği gibi birleştirilmiş verinin ortak standart sapmasını (*pooled standard deviation*) da ifade edebilir çünkü iki grubun standart sapması eşit olduğunda ortak standart sapma gruplardan birinin standart sapmasına eşit olur (varsayılan ayarlarda ikinci grubun standart sapması birinci grubun standart sapmasına eşittir). Elde edilen sonuçlara göre, her bir grupta 50'şer katılımcı ile aradaki iki birimlik fark %43.1 güç oranı ile tespit edilebilir. Şayet, daha fazla katılımcı seçme olanağı var ise, her bir grup için 142 katılımcı ile iki birimlik fark %80 güç ile tespit edilebilir.

Çift yönlü hipotez testi: Kitledeki bir grubun ortalamasının (μ_1) diğer grubun (μ_2) ortalamasına eşit olmadığı düşünülüyorsa, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_1$

Test istatistiği Denklem 50 ile aynıdır. Farklı olarak, kestirilecek $\hat{\mu}_1$ değeri, $\hat{\mu}_2$ değerinin hem solunda hem de sağında olma ihmali göz önünde bulundurulduğu için çift kuyruk hipotez testi gerçekleştirilir. Bundan dolayı, α yerine $\alpha/2$ kullanılır. İstatistiksel güç

$$t_k = \Phi_t^{-1}(\alpha/2, v; 0) \tag{56}$$

$$1 - \beta = 1 - \Phi_t(t_k, v; t) + \Phi_t(-t_k, v; t) \tag{57}$$

olarak, örneklem büyüklüğü ise

$$t_{\alpha/2} + t_\beta = \Phi_z^{-1}(\alpha/2, v; 0) + \Phi_z^{-1}(\beta, v; 0) \tag{58}$$

$$n_2 = \left(t_{\alpha/2} + t_\beta\right)^2 \left(\frac{\hat{\sigma}_1^{\,2}}{\kappa} + \hat{\sigma}_2^{\,2} \over (\hat{\mu}_1 - \hat{\mu}_2)^2 \right) \tag{59}$$

şeklinde hesaplanır.

Örnek: Bir araştırmacı, kızların depresyon seviyelerinin erkeklere eşit olup olmadığını bulmaya çalışmaktadır. Burada en küçük anlamlı farkın 2 birim olduğu farz edilsin (kızlar için 26 erkekler için ise 24). Her bir grupta 50'şer katılımcı ile elde edilecek istatistiksel güç ya da %80 güç için örneklem büyüklüğü `alternative="not equal"` olacak şekilde tanımlanarak

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, kappa = 1,
               alpha = 0.05, n = 50, alternative = "not equal")
# Difference between two means
# (independent samples t test)
# H0: mu1 = mu2
# HA: mu1 != mu2
# ------------------------------
#  Statistical power = 0.311
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "not equal"
# Degrees of freedom = 98
# Non-centrality parameter = 1.481
# Type I error rate = 0.05
# Type II error rate = 0.689

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, kappa = 1,
               alpha = 0.05, power = 0.8, alternative = "not equal")
# Difference between two means
# (independent samples t test)
```

```
# H0: mu1 = mu2
# HA: mu1 != mu2
# ------------------------------
#  Statistical power = 0.8
#  n1 = 180
#  n2 = 180
# ------------------------------
# Alternative = "not equal"
# Degrees of freedom = 357.56
# Non-centrality parameter = 2.809
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Hesaplamalara göre, her bir grupta 50'şer katılımcı ile aradaki iki birimlik fark %31.1 güç oranı ile tespit edilebilir. Şayet, daha fazla katılımcı seçme imkânı var ise, hipotez testini %80 güç oranı ile kestirmek için her bir grupta en az 179 katılımcıya ihtiyaç olduğu anlaşılmaktadır.

Aşağı olmama (non-inferiority) ya da üstünlük (superiority) hipotez testi: Aşağı olmama (*non-inferiority*) hipotez testinde, kitledeki iki grubun ortalamalarının farkı ($\mu_1 - \mu_2$) 0'ın solundaki sınır değerden (δ negatif) daha büyük olduğu, ya da 0'ın sağındaki sınır değerinden (δ pozitif) daha küçük olduğu düşünülür. Değişkenin yüksek değerlerinin olumlu bir olguyu ifade ettiği durumlarda (örn. başarı testi puanı) sınır değeri genellikle negatiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\mu_1 - \mu_2 \leq \delta$

H_A: $\mu_1 - \mu_2 > \delta$

ya da değişkenin yüksek değerlerinin olumsuz bir olguyu ifade ettiği durumlarda (örn. depresyon puanı) sınır değeri genellikle pozitiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\mu_1 - \mu_2 \geq \delta$

H_A: $\mu_1 - \mu_2 < \delta$

Üstünlük (*superiority*) hipotez testinde ise kitledeki iki grubun ortalamaları farkı ($\mu_1 - \mu_2$) 0'ın solundaki sınır değerinden (δ negatif) daha küçük olduğu, ya da 0'ın sağındaki sınır değerinden (δ pozitif) daha büyük olduğu düşünülür. Değişkenin yüksek değerlerinin olumlu bir olguyu ifade ettiği durumlarda (örn. başarı testi puanı) sınır değeri genellikle pozitiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\mu_1 - \mu_2 \leq \delta$

H_A: $\mu_1 - \mu_2 > \delta$

ya da değişkenin yüksek değerlerinin olumsuz bir olguyu ifade ettiği durumlarda (örn. depresyon puanı) sınır değeri genellikle negatiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\mu_1 - \mu_2 \geq \delta$

H_A: $\mu_1 - \mu_2 < \delta$

Hem aşağı olmama (*non-inferiority*) hem de üstünlük (*superiority*) testleri için test istatistiği

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2 - \delta}{SH(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{\hat{\mu}_1 - \hat{\mu}_2 - \delta}{\sqrt{\frac{\hat{\sigma}_1^{\,2}}{n_1} + \frac{\hat{\sigma}_2^{\,2}}{n_2}}} \tag{60}$$

şeklinde hesaplanır. İstatiksel gücü hesaplamak için Denklem 52 ve 53 kullanılır. Denklem 54'teki işlem uygulandıktan sonra, ikinci grup için örneklem büyüklüğü

$$n_2 = \left(t_\alpha + t_\beta\right)^2 \left(\frac{\frac{\hat{\sigma}_1^{\,2}}{\kappa} + \hat{\sigma}_2^{\,2}}{(\hat{\mu}_1 - \hat{\mu}_2 - \delta)^2} \right) \tag{61}$$

şeklinde hesaplanır. İkinci grup için örneklem büyüklüğü $n_1 = \kappa n_2$ eşitliğinden elde edilir.

Örnek: Bir araştırmacı, COVID-19 döneminde ortaokullardaki kız öğrencilerin psikolojik sağlamlık seviyelerinin erkek katılımcılara kıyasla daha yüksek olduğunu öne sürmektedir. İki ortalama arasındaki fark δ = -1 olsa bile pratik anlamda μ_1'in μ_2'den daha az olmadığı farz edilsin. Her bir gruptan 50 katılımcı olacak şekilde istatistiksel güç ya da %80 güç için örneklem büyüklüğü `margin=-1` ve `alternative="non-inferior"` olacak şekilde tanımlanarak

```
pwrss.t.2means(mu1 = 50, mu2 = 48, sd1 = 7.59, margin = -1,
               alpha = 0.05, n = 50, alternative = "non-inferior")
# Difference between two means
# (independent samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.625
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "non-inferior"
# Degrees of freedom = 98
# Non-centrality parameter = 1.976
# Type I error rate = 0.05
# Type II error rate = 0.375
```

```
pwrss.t.2means(mu1 = 50, mu2 = 48, sd1 = 7.59, margin = -1,
               alpha = 0.05, power = 0.8, alternative = "non-inferior")
# Difference between two means
# (independent samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.8
#  n1 = 80
#  n2 = 80
# ------------------------------
# Alternative = "non-inferior"
# Degrees of freedom = 157.82
# Non-centrality parameter = 2.498
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Görüldüğü gibi, her bir grupta 50'şer katılımcı ile *non-inferiority* hipotez testi %62.5 güç oranı ile gerçekleştirilebilir. Şayet daha fazla katılımcı seçmek mümkün ise %80 güç oranı için her bir grupta an ez 80 katılımcı gerekmektedir.

Şimdi de, kız öğrencilerin psikolojik sağlamlık seviyelerinin erkek öğrencilere kıyasen pratik anlamda daha yüksek olduğu öne sürülsün. İki ortalama arasındaki fark δ = 1'den büyük olmalı ki pratik anlamda μ_1'in μ_2'den daha yüksek olduğu kabul edilsin. Yine her bir gruptan 50'şer katılımcı olacak şekilde istatistiksel güç ya da %80 güç için örneklem büyüklüğü `alternative="superior"` olacak şekilde tanımlanarak

```
pwrss.t.2means(mu1 = 50, mu2 = 48, sd1 = 7.59, margin = 1,
               alpha = 0.05, n = 50, alternative = "superior")
# Difference between two means (independent samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.161
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "superior"
# Degrees of freedom = 98
# Non-centrality parameter = 0.659
# Type I error rate = 0.05
# Type II error rate = 0.839
```

```
pwrss.t.2means(mu1 = 50, mu2 = 48, sd1 = 7.59, margin = 1,
               alpha = 0.05, power = 0.80, alternative = "superior")
# Difference between two means (independent samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.8
#  n1 = 714
#  n2 = 714
# ------------------------------
# Alternative = "superior"
# Degrees of freedom = 1424.16
# Non-centrality parameter = 2.488
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Her bir gruptan 50'şer katılımcı ile elde edilen istatistiksel güç oranı %16.1'dir. Şayet daha fazla katılımcı seçmek mümkün ise %80 güç oranı için her bir grupta en az 713 katılımcı gereklidir.

Eşdeğerlik (equivalence) hipotez testi: Eşdeğerlik hipotez testinde, kitledeki iki grubun ortalamaları farkının mutlak değerinin ($|\mu_1 - \mu_2|$) sınır değerden (δ her zaman pozitif) daha küçük olduğu düşünülür. Yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $|\mu_1 - \mu_2| \geq \delta$

H_A: $|\mu_1 - \mu_2| < \delta$

Test istatistiği

$$t = \frac{|\hat{\mu}_1 - \hat{\mu}_2| - \delta}{SH(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{|\hat{\mu}_1 - \hat{\mu}_2| - \delta}{\sqrt{\frac{\hat{\sigma}_1^{\,2}}{n_1} + \frac{\hat{\sigma}_2^{\,2}}{n_2}}} \tag{62}$$

şeklinde hesaplanır. İstatiksel güç

$$t_k = \Phi_t^{-1}(\alpha, v;\ 0) \tag{63}$$

$$1 - \beta = 2\big(1 - \Phi_t(t_k, v; t)\big) - 1 \tag{64}$$

olarak, örneklem büyüklüğü ise

$$t_\alpha + t_{\beta/2} = \Phi_t^{-1}(\alpha, v; 0) + \Phi_t^{-1}(\beta/2, v; 0) \tag{65}$$

$$n_2 = \left(t_\alpha + t_{\beta/2}\right)^2 \left(\frac{\frac{\hat{\sigma}_1^{\,2}}{\kappa} + \hat{\sigma}_2^{\,2}}{(|\hat{\mu}_1 - \hat{\mu}_2| - \delta)^2} \right) \quad (66)$$

şeklinde hesaplanır.

Örnek: Bir araştırmacı, kız öğrencilerin psikolojik sağlamlık seviyelerinin erkek öğrencilerinkine eşdeğer olduğunu öne sürmektedir. İki ortalama arasındaki fark 1'den (δ = 1) küçük olduğu sürece pratik anlamda μ_1'in μ_2'e eşdeğer olduğu kabul edilsin. Yine her bir grupta 50'şer katılımcı olacak şekilde istatistiksel güç ya da %80 güç için örneklem büyüklüğü `margin=1` ve `alternative="equivalent"` olacak şekilde tanımlanarak

```
pwrss.t.2means(mu1 = 50, mu2 = 48, sd1 = 7.59, margin = 1,
               alpha = 0.05, n2 = 50, alternative = "equivalent")
# Error: design is not feasible

pwrss.t.2means(mu1 = 50, mu2 = 48, sd1 = 7.59, margin = 1,
              alpha = 0.05, power = 0.80, alternative = "equivalent")
# Difference between two means (independent samples t test)
# H0: |mu1 - mu2| >= margin
# HA: |mu1 - mu2| < margin
# ------------------------------
#  Statistical power = 0.8
#  n1 = 988
#  n2 = 988
# ------------------------------
# Alternative = "equivalent"
# Degrees of freedom = 1973
# Non-centrality parameter = 2.928
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Buna göre, her bir grupta 50'şer katılımcının olması güç oranını hesaplamak için yeterli değildir. Bu durum verilen hatadan görülmektedir (`Error: design is not feasible`). Öte yandan, %80 güç ile eşdeğerlik test edilmek isteniyorsa her bir grupta en az 988'er kişi gerekmektedir.

Bağımlı örneklemler *t* testi: Bir grup için birinci zaman dilimindeki ölçülen rassal bir X_1 değişkenin ortalama değeri μ_1 ve varyansı σ_1^2, ikinci zaman diliminde tekrar ölçülen bu değişkeninin (X_2) ortalama değeri μ_2 ve varyansı σ_2^2 olarak kabul edilsin. X_1 ve X_2 ölçümleri aynı bireylere ait olduğu için bu iki değişken arasında ilişki bulunmaktadır. Büyüklüğü n olan bir örneklemde birinci zaman diliminde X_1 ile ilgili $x_{11}, x_{12}, x_{13}, \ldots, x_{1n}$ değerleri, ikinci zaman diliminde ise X_2 ile ilgili $x_{21}, x_{22}, x_{23}, \ldots, x_{2n}$ değerleri gözlensin. X_1 ve X_2'in ortalamalarının tahmin edicileri

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i} \tag{67}$$

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{2i} \tag{68}$$

varyanslarının tahmin edicileri ise

$$\hat{\sigma}_1{}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_{1i} - \hat{\mu}_1)^2 \tag{69}$$

$$\hat{\sigma}_2{}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_{2i} - \hat{\mu}_2)^2 \tag{70}$$

olarak ifade edilir. $\hat{\mu}_1 - \hat{\mu}_2$ farkının standart hatası ise

$$SH(\hat{\mu}_1 - \hat{\mu}_2) = \sqrt{\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{n}} \tag{71}$$

şeklinde hesaplanır. Burada r_{12}, X_1 ve X_2 arasındaki korelasyon olup

$$r_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1\hat{\sigma}_2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_{1i} - \hat{\mu}_1)(x_{2i} - \hat{\mu}_2)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{1i} - \hat{\mu}_1)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{2i} - \hat{\mu}_2)^2}} \tag{72}$$

şeklinde tanımlanır.

Tek yönlü hipotez testi: Kitlede bir değişkenin birinci zaman dilimindeki ortalamasının (μ_1) ikinci zaman dilimindeki ortalamasından (μ_2) daha küçük ya da daha büyük olduğu düşünülüyorsa, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\mu_1 \geq \mu_2$ (ya da $\mu_1 \leq \mu_2$)

H_A: $\mu_1 < \mu_2$ (ya da $\mu_1 > \mu_2$)

Test istatistiği

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{SH(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{n}}} \tag{73}$$

şeklinde tanımlanır. İstatiksel güç, serbestlik derecesi $v = n - 1$ tanımlanarak Denklem 52 ve 53'de olduğu gibi hesaplanır. Denklem 56'daki işlem yapıldıktan sonra, en küçük gerekli örneklem büyüklüğü

$$n = (t_\alpha + t_\beta)^2 \left(\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{(\hat{\mu}_1 - \hat{\mu}_2)^2} \right) \quad (74)$$

şeklinde hesaplanır. Standardize ortalama farkı olarak Cohen d kullanılacaksa, istatistiksel güç ve örneklem büyüklüğü hesaplamalarında $\hat{\mu}_1 = d$, $\hat{\mu}_2 = 0$, $\hat{\sigma}_1{}^2 = \sqrt{1/2(1 - r_{12})}$ ve $\hat{\sigma}_2{}^2 = \sqrt{1/2(1 - r_{12})}$ tanımlamaları yapılır.

Örnek: Bir araştırmacı, COVID-19 döneminde üniversitelerde okuyan öğrencilerin depresyon seviyelerini düşürmek için bilişsel davranışsal psikoterapiye dayanan bir haftalık program düzenlemeyi düşünmektedir. Araştırmacı, bu programın başında ve sonunda katılımcıların depresyon seviyeleri ölçecektir (ön test ve son test). Bulus ve Koyuncu (2021) psikolojik danışmanlık alanında bilişsel olmayan çıktılar için ön testin son testteki varyansın 0.29'unu açıkladığını raporlamışlardır ($r_{12}^2 = 0.29$). Bu determinasyon katsayısı ön test ve son test arasındaki korelasyonun $r_{12} = \sqrt{0.29} = 0.54$ olduğunu göstermektedir. Ayrıca, daha önceki çalışmalardan depresyon puanlarının standart sapmasının 6.75 civarında olduğunu bilinmektedir (Hisli, 1989). Depresyon belirtilerinde en küçük anlamlı azalmanın 2 birim olduğu farz edilirse (ön test ortalaması 26 ve son test ortalaması 24 puan), 50 katılımcı ile elde edilecek istatistiksel güç ya da %80 güç için örneklem büyüklüğü `paired=TRUE` ve `alternative="greater"` olacak şekilde tanımlanarak

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, n = 50,
               paired = TRUE, paired.r = 0.54, alternative = "greater")
# Difference between two means
# (paired samples t test)
# H0: mu1 = mu2
# HA: mu1 > mu2
# ------------------------------
#  Statistical power = 0.695
#  n = 50
# ------------------------------
# Alternative = "greater"
# Degrees of freedom = 49
# Non-centrality parameter = 2.184
# Type I error rate = 0.05
# Type II error rate = 0.305

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, power = 0.8,
               paired = TRUE, paired.r = 0.54, alternative = "greater")
# Difference between two means
```

```
# (paired samples t test)
# H0: mu1 = mu2
# HA: mu1 > mu2
# ------------------------------
#  Statistical power = 0.8
#  n = 67
# ------------------------------
# Alternative = "greater"
# Degrees of freedom = 65.31
# Non-centrality parameter = 2.516
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Daha önce belirtildiği üzere, `sd1` argümanı birinci grubun standart sapmasını temsil ettiği gibi birleştirilmiş verinin ortak standart sapmasını (*pooled standard deviation*) da ifade edebilir çünkü varsayılan ayarlarda ikinci grubun standart sapması birinci grubun standart sapmasına eşitlenmiştir. Buradaki hesaplamaya göre, 50 katılımcı ile iki birimlik fark %69.5 güç oranı ile tespit edilebilir. Ayrıca, şayet daha fazla katılımcı seçmek mümkün ise, %80 güç oranı için en az 67 katılımcıya ihtiyaç duyulduğu görülmektedir. Depresyon seviyesinde bir artış beklenseydi, yani ön test puanları ortalamasının son test puanları ortalamasından daha düşük olduğunu bulunmaya çalışılsaydı `alternative="less"` argümanını kullanılacaktı.

Çift yönlü hipotez testi: Kitlede bir değişkenin birinci zaman dilimindeki ortalamasının (μ_1) ikinci zaman dilimindeki ortalamasından (μ_2) farklı olduğu düşünülüyorsa, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_1$

Test istatistiği Denklem 73'teki ile benzerdir. Farklı olarak, kestirilecek $\hat{\mu}_1$ değeri, $\hat{\mu}_2$ değerinin hem solunda hem de sağında olma ihmali göz önünde bulundurulduğu için çift kuyruk hipotez testi gerçekleştirilir. Bundan dolayı, α yerine $\alpha/2$ kullanılır. İstatiksel güç, serbestlik derecesi $v = n - 1$ tanımlanarak Denklem 56 ve 57'de olduğu gibi hesaplanır. Denklem 54'te α yerine $\alpha/2$ ve $v = n - 1$ tanımlanarak işlem yapıldıktan sonra, örneklem büyüklüğü

$$n = \left(t_{\alpha/2} + t_{\beta}\right)^2 \left(\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{(\hat{\mu}_1 - \hat{\mu}_2)^2}\right) \tag{75}$$

şeklinde hesaplanır.

Örnek: Bir önceki örnek bağlamında, depresyon seviyesinde bir düşüş olma ihtimali olabileceği gibi bir yükselme ihtimali de söz konusu olduğunda, 50 katılımcı ile elde edilecek istatistiksel güç ya da %80 güç için örneklem büyüklüğü `alternative="not equal"` olacak şekilde tanımlanarak

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, n = 50,
               paired = TRUE, paired.r = 0.54, alternative = "not equal")
# Difference between two means
# (paired samples t test)
# H0: mu1 = mu2
# HA: mu1 != mu2
# ------------------------------
#  Statistical power = 0.572
#  n = 50
# ------------------------------
# Alternative = "not equal"
# Degrees of freedom = 49
# Non-centrality parameter = 2.184
# Type I error rate = 0.05
# Type II error rate = 0.428

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, power = 0.8,
               paired = TRUE, paired.r = 0.54, alternative = "not equal")
# Difference between two means
# (paired samples t test)
# H0: mu1 = mu2
# HA: mu1 != mu2
# ------------------------------
#  Statistical power = 0.8
#  n = 85
# ------------------------------
# Alternative = "not equal"
# Degrees of freedom = 83.21
# Non-centrality parameter = 2.835
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Buna göre, 50 katılımcı ile aradaki iki birimlik fark %57.2 güç oranı ile tespit edilebilir. Şayet, daha fazla katılımcı seçme olanağı var ise, %80 güç oranı için en az 85 katılımcıya ihtiyaç vardır.

Aşağı olmama (non-inferiority) ya da üstünlük (superiority) hipotez testi: Aşağı olmama (*non-inferiority*) hipotez testinde bir değişkenin birinci zaman ve ikinci zaman dilimindeki ortalama farkın ($\mu_1 - \mu_2$) 0'ın solundaki sınır değerinden (δ negatif) daha büyük olduğu ya da 0'ın sağındaki sınır değerinden (δ pozitif) daha küçük olduğu düşünüldüğünde kullanılır. Değişkenin yüksek değerlerinin olumlu bir olguyu ifade ettiği durumlarda (örn. başarı testi puanı) sınır değeri genellikle negatiftir yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\mu_1 - \mu_2 \leq \delta$

$H_A: \mu_1 - \mu_2 > \delta$

ya da değişkenin yüksek değerlerinin olumsuz (örn. depresyon puanı) bir olguyu ifade ettiği durumlarda sınır değeri genellikle pozitiftir yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

$H_0: \mu_1 - \mu_2 \geq \delta$

$H_A: \mu_1 - \mu_2 < \delta$

Üstünlük (*superiority*) hipotez testinde ise bu farkın ($\mu_1 - \mu_2$) 0'ın solundaki sınır değerinden (δ negatif) daha küçük olduğu, ya da 0'ın sağındaki sınır değerinden (δ pozitif) daha büyük olduğu düşünülür. Değişkenin yüksek değerlerinin olumlu bir olguyu ifade ettiği durumlarda sınır değeri genellikle pozitiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

$H_0: \mu_1 - \mu_2 \leq \delta$

$H_A: \mu_1 - \mu_2 > \delta$

ya da değişkenin yüksek değerlerinin olumsuz bir olguyu ifade ettiği durumlarda sınır değeri genellikle negatiftir ve yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

$H_0: \mu_1 - \mu_2 \geq \delta$

$H_A: \mu_1 - \mu_2 < \delta$

Hem aşağı olmama (*non-inferiority*) hem de üstünlük (*superiority*) hipotez testleri için test istatistiği

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2 - \delta}{SH(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{\hat{\mu}_1 - \hat{\mu}_2 - \delta}{\sqrt{\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{n}}} \quad (76)$$

şeklinde hesaplanır. İstatiksel güç, serbestlik derecesi $v = n - 1$ tanımlanarak Denklem 52 ve 53'te olduğu gibi hesaplanır. Denklem 54'te $v = n - 1$ tanımlanıp işlem yapıldıktan sonra, örneklem büyüklüğü ise

$$n = (t_\alpha + t_\beta)^2 \left(\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{(\hat{\mu}_1 - \hat{\mu}_2 - \delta)^2} \right) \quad (77)$$

şeklinde hesaplanır.

Örnek: Bir önceki örnek tekrardan ele alınsın fakat bu sefer hipotez testi şu şekilde ifade edilsin: Ön test ile son test arasındaki fark -1'den büyük olduğu sürece depresyon düzeyinde bir artışın olmadığı farz edilsin (sınır değer $\delta = -1$). Başka bir deyişle, ön testin son testten büyük olduğu ve ön testin son testten en fazla 1 puan küçük olması durumunda depresyon düzeyinde artış olmadığı düşünülsün. Elli

katılımcı ile elde edilecek istatistiksel güç ya da %80 güç için örneklem büyüklüğü `margin=-1` ve `alternative="non-inferior"` olacak şekilde tanımlanarak

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, n = 50,
               paired = TRUE, paired.r = 0.54,
               alternative = "non-inferior", margin = -1)
# Difference between two means
# (paired samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.944
#  n = 50
# ------------------------------
# Alternative = "non-inferior"
# Degrees of freedom = 49
# Non-centrality parameter = 3.276
# Type I error rate = 0.05
# Type II error rate = 0.056

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, power = 0.8,
               paired = TRUE, paired.r = 0.54,
               alternative = "non-inferior", margin = -1)
# Difference between two means
# (paired samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.8
#  n = 31
# ------------------------------
# Alternative = "non-inferior"
# Degrees of freedom = 29.34
# Non-centrality parameter = 2.552
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Buna göre, sınır değeri -1 olacak şekilde, 50 katılımcı ile ön test ve son test ortalamaları arasındaki iki birimlik fark %94.4 güç oranı ile tespit edilebilir. Ayrıca, %80 güç oranı için 31 katılımcının yeterli olduğu hesaplanmıştır.

Şimdi de, ön test ile son test arasındaki fark 1'den büyük olduğu sürece depresyon düzeyinde bir azalma meydana geldiği farz edilsin (sınır değer $\delta = 1$). Elli katılımcı ile elde edilecek istatistiksel

güç ya da %80 güç için örneklem büyüklüğü `margin=1` ve `alternative="superior"` olacak şekilde tanımlanarak

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, n = 50,
               paired = TRUE, paired.r = 0.54,
               alternative = "superior", margin = 1)
# Difference between two means
# (paired samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.285
#  n = 50
# ------------------------------
# Alternative = "superior"
# Degrees of freedom = 49
# Non-centrality parameter = 1.092
# Type I error rate = 0.05
# Type II error rate = 0.715

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, power = 0.8,
               paired = TRUE, paired.r = 0.54,
               alternative = "superior", margin = 1)
# Difference between two means
# (paired samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.8
#  n = 261
# ------------------------------
# Alternative = "superior"
# Degrees of freedom = 259.67
# Non-centrality parameter = 2.494
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Buna göre, sınır değeri 1 olacak şekilde, elli katılımcı ile ön test ve son test ortalamaları arasındaki iki birimlik fark %28.5 güç oranı ile tespit edilebilir. Şayet daha fazla katılımcı seçme olanağı var ise, %80 güç oranı için en az 261 katılımcıya ihtiyaç vardır.

Eşdeğerlik (equivalence) hipotez testi: Eşdeğerlik hipotez testinde, ölçümlerin ortalama farkının ($\mu_1 - \mu_2$) 0'ın solundaki sınır değerinden (δ negatif) daha büyük olduğu, ya da 0'ın sağındaki sınır

değerinden (δ pozitif) daha küçük olduğu düşünülür. Bu durumda, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $|\mu_1 - \mu_2| \geq \delta$

H_A: $|\mu_1 - \mu_2| < \delta$

Test istatistiği

$$t = \frac{|\hat{\mu}_1 - \hat{\mu}_2| - \delta}{SH(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{|\hat{\mu}_1 - \hat{\mu}_2| - \delta}{\sqrt{\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{n}}} \tag{78}$$

şeklinde elde edilir. İstatiksel güç Denklem 63 ve 64'te serbestlik derecesi $v = n - 1$ tanımlanarak hesaplanır. Denklem 65'te serbestlik derecesi $v = n - 1$ olarak tanımlanıp gerekli düzenlemeler yapıldıktan sonra örneklem büyüklüğü

$$n = \left(t_\alpha + t_{\beta/2}\right)^2 \left(\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{(|\hat{\mu}_1 - \hat{\mu}_2| - \delta)^2}\right) \tag{79}$$

şeklinde hesaplanır.

Örnek: Bir önceki örnek bağlamında, ön test ile son test arasındaki farkın mutlak değeri 1'den küçük olduğu sürece depresyon düzeyinde bir değişme meydana gelmediği farz edilsin (sınır değer $\delta = 1$). Elli katılımcı ile elde edilen istatistiksel güç ya da %80 güç için örneklem büyüklüğü `margin=1` ve `alternative="equivalent"` olacak şekilde tanımlanarak

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, n = 50,
               paired = TRUE, paired.r = 0.54,
               alternative = "equivalent", margin = 1)
# Error: design is not feasible

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, power = 0.8,
               paired = TRUE, paired.r = 0.54,
               alternative = "equivalent", margin = 1)
# Difference between two means
# (paired samples t test)
# H0: |mu1 - mu2| >= margin
# HA: |mu1 - mu2| < margin
# ------------------------------
#  Statistical power = 0.8
#  n = 361
# ------------------------------
# Alternative = "equivalent"
# Degrees of freedom = 359.59
# Non-centrality parameter = 2.933
```

```
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Burada 50 kişilik örneklem büyüklüğünün güç oranını hesaplamak için yeterli olmadığı verilen hatadan görülmektedir (`Error: design is not feasible`). Ayrıca, %80 güç oranı ile eşdeğerlik tespit edilecekse en az 361 kişi gerektiği hesaplanmıştır.

Tek Bir Korelasyonunun Sabitle Karşılaştırılması

İki değişken arasındaki Pearson korelasyonunun ($\hat{r}$) belirli sabit bir değer (r_0) ile karşılaştırılması, Fisher dönüşümü uygulandıktan sonra z testi ile yapılır (Cohen, 1988). Fisher dönüşümleri aşağıdaki gibi elde edilir:

$$\hat{z} = \frac{1}{2}\log\left(\frac{1+\hat{r}}{1-\hat{r}}\right) \tag{80}$$

$$z_0 = \frac{1}{2}\log\left(\frac{1+r_0}{1-r_0}\right) \tag{81}$$

Test istatistiği ise

$$z = \frac{\hat{z} - z_0}{\sqrt{\frac{1}{n-3}}} \tag{82}$$

şeklinde ifade edilir. Burada n örneklem büyüklüğüdür.

Tek yönlü hipotez testi: Tek yönlü hipotez testi, kitlede iki değişken arasındaki r korelasyonunun sabit bir değerden (r_0) daha küçük ya da daha büyük olduğu düşünülüyorsa kullanılır. Yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $r \geq r_0$ (ya da $r \leq r_0$)

H_A: $r < r_0$ (ya da $r > r_0$)

İstatistiksel güç Denklem 7 ve 8 kullanılarak hesaplanır. Denklem 9'daki işlem yapıldıktan sonra, örneklem büyüklüğü

$$n = \frac{\left(z_\alpha + z_\beta\right)^2}{(\hat{z} - z_0)^2} + 3 \tag{83}$$

şeklinde hesaplanır. `pwrss` R paketinde $r_0 = 0$ varsayılan ayardır. Böylece, korelasyonun sıfırdan büyük, küçük veya farklı olduğunu bulmaya çalışan bir araştırmacı genellikle kodlarda bu tanımlamayı yapmaya gerek duymaz.

Örnek: Bir araştırmacı, işbirlikçi öğrenme ortamı ve okul aidiyeti arasında pozitif bir ilişkinin olduğunu öne sürmektedir. Daha önce yapılan bir araştırmada, işbirlikçi öğrenme ortamı ile okul aidiyeti arasında 0.24 büyüklüğünde pozitif bir ilişki olduğu bulunmuştur (Özcan ve Buluş, 2022). Beklenen korelasyon değeri hakkında herhangi bir bilgi mevcut değilse Cohen (1988) ya da Gignac ve Szodorai (2016) sınıflamaları kullanılabilir. Elli katılımcı ile istatistiksel güç ya da %80 güç için örneklem büyüklüğü `alternative="greater"` olacak şekilde tanımlanarak

```
pwrss.z.corr(r = .24, n = 50, alternative = "greater")
# One correlation compared to a constant (one sample z test)
# H0: r = r0
# HA: r > r0
# ------------------------------
#  Statistical power = 0.513
#  n = 50
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 1.678
# Type I error rate = 0.05
# Type II error rate = 0.487

pwrss.z.corr(r = .24, power = .8, alternative = "greater")
# One correlation compared to a constant (one sample z test)
# H0: r = r0
# HA: r > r0
# ------------------------------
#  Statistical power = 0.8
#  n = 107
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır. Sonuçlara göre, orta düzeyde kabul edilebilecek $r = 0.24$ korelasyon katsayısının 0'dan büyük olduğunu bulmaya çalışan bir araştırmacı 50 kişi ile hipotez testini %51.3 güç oranı ile test edebilir. Diğer taraftan, çalışmaya daha fazla katılımcı alma olanağı varsa %80 güç oranı ile hipotez testini gerçekleştirmek için en az 107 katılımcıya gerek olduğu görülmektedir.

Çift yönlü hipotez testi: Kitledeki r korelasyonunun sabit bir korelasyon değerinden (r_0) farklı olduğu düşünülüyorsa, yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $r = r_0$

H_A: $r \neq r_0$

İstatistiksel güç Denklem 11 ve 12'teki gibi hesaplanır. Denklem 9'da α yerine $\alpha/2$ konularak gerekli düzenlemeler yapıldıktan sonra örneklem büyüklüğü

$$n = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2}{(\hat{z} - z_0)^2} + 3 \tag{84}$$

denkleminden elde edilir.

Örnek: Bir önceki örnek bağlamında, şayet araştırmacı işbirlikçi öğrenme ortamı ve okul aidiyeti arasındaki ilişkinin 0'dan farklı olduğunu kanıtlamaya çalışıyorsa `alternative="not equal"` olacak şekilde tanımlama yapmalıdır.

Bu durumda hesaplamalar

```
pwrss.z.corr(r = .24, n = 50, alternative = "not equal")
# One correlation compared to a constant
# (one sample z test)
# H0: r = r0
# HA: r != r0
# ------------------------------
#  Statistical power = 0.389
#  n = 50
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 1.678
# Type I error rate = 0.05
# Type II error rate = 0.611

pwrss.z.corr(r = .24, power = .8, alternative = "not equal")
# One correlation compared to a constant
# (one sample z test)
# H0: r = r0
# HA: r != r0
# ------------------------------
#  Statistical power = 0.8
#  n = 135
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 2.802
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde yapılır. Elde edilen sonuçlara göre, orta düzeyde kabul edilebilecek $r = 0.24$ korelasyon katsayısının 0'dan farklı olduğunu bulmaya çalışan bir araştırmacı 50 kişi ile hipotez testini %38.9 güç

oranı ile test edebilir. Ayrıca, çalışmaya daha fazla katılımcı alma olanağı varsa %80 güç oranı ile hipotez testini gerçekleştirmek için en az 135 katılımcıya gerek vardır.

İki Korelasyon Farkı

Büyüklüğü n_1 olan örneklemde iki değişken arasındaki Pearson korelasyonu $\hat{r}_1$ ile, büyüklüğü n_2 olan örneklemde iki değişken arasındaki Pearson korelasyonu $\hat{r}_2$ ile gösterilsin. $\hat{r}_1$ ve $\hat{r}_2$ korelasyon katsayıları arasındaki fark, Fisher dönüşümü uygulanarak, z testi ile test edilir (Cohen, 1988). Buna göre

$$\hat{z}_1 = \frac{1}{2}\log\left(\frac{1+\hat{r}_1}{1-\hat{r}_1}\right) \tag{85}$$

$$\hat{z}_2 = \frac{1}{2}\log\left(\frac{1+\hat{r}_2}{1-\hat{r}_2}\right) \tag{86}$$

olarak tanımlandığında, test istatistiği

$$z = \frac{\hat{z}_1 - \hat{z}_2}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}} \tag{87}$$

şeklinde ifade edilir.

Tek yönlü hipotez testi: Kitlede birbirinden bağımsız iki Pearson korelasyonu değerinden birinin (r_1) diğerinden (r_2) daha küçük ya da daha büyük olduğu düşünüldüğünde kullanılır. Yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $r_1 \geq r_2$ (ya da $r_1 \leq r_2$)

H_A: $r_1 < r_2$ (ya da $r_1 > r_2$)

İstatistiksel güç Denklem 7 ve 8 kullanılarak hesaplanır. Denklem 9'daki işlem gerçekleştirdikten sonra, ikinci grup için örneklem büyüklüğü

$$f(n_2) = \left(\frac{1}{\kappa n_2 - 3} + \frac{1}{n_2 - 3}\right) - \frac{(z_1 - z_2)^2}{(z_\alpha + z_\beta)^2} = 0 \tag{88}$$

denkleminden R programındaki `uniroot()` kök bulma algoritması kullanılarak elde edilir. $\kappa = n_1/n_2$ olduğu daha önce belirtilmişti; o halde birinci grup için $n_1 = \kappa n_2$ eşitliğinden bulunur.

Örnek: Bireyci toplumlarda işbirlikçi öğrenme ortamı ve okul aidiyeti arasında $r_1 = 0.23$ büyüklüğünde pozitif bir ilişkinin olduğunu ve kolektivist toplumlarda işbirlikçi öğrenme ortamı ve okul aidiyeti arasında $r_2 = 0.25$ büyüklüğünde pozitif bir ilişkinin olduğu daha önceki çalışmalardan bilinmektedir (Özcan ve Buluş, 2022). Her bir gruptan 50'şer kişilik bir örneklem ile istatistiksel güç ya da %80 güç için örneklem büyüklüğü `alternative="less"` olacak şekilde tanımlanarak

```
pwrss.z.2corrs(r1 = .23, r2 = 0.25,
               n2 = 50, alternative = "less")
# Difference between two correlations
# (independent samples z test)
# H0: r1 = r2
# HA: r1 < r2
# ------------------------------
#  Statistical power = 0.062
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "less"
# Non-centrality parameter = -0.103
# Type I error rate = 0.05
# Type II error rate = 0.938

pwrss.z.2corrs(r1 = .23, r2 = 0.25,
               power = 0.8, alternative = "less")
# Difference between two correlations
# (independent samples z test)
# H0: r1=r2
# HA: r1 < r2
# ------------------------------
#  Statistical power = 0.8
#  n1 = 27455
#  n2 = 27455
# ------------------------------
# Alternative = "less"
# Non-centrality parameter = -2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanabilir. Sonuçlara göre, bireyci toplumlardaki korelasyonun ($r_1 = 0.23$) kolektivist toplumlardaki ($r_1 = 0.25$) korelasyondan daha düşük olduğunu ve 0.02 birimlik bir farkın pratikte anlamlı olduğunu iddia eden bir araştırmacı, her bir grupta 50'şer kişi ile hipotez testini %6.2 güç oranı ile test edebilir. Hipotez testini %80 güç oranı ile gerçekleştirmek için her bir gruptan en az 27455 katılımcıya ihtiyaç vardır.

Çift yönlü hipotez testi: Kitlede birbirinden bağımsız iki Pearson korelasyonu değerinden birinin (r_1) diğerine (r_2) eşit olmadığı düşünüldüğünde kullanılır. Yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

$H_0: r_1 = r_2$

$H_A: r_1 \neq r_2$

İstatistiksel güç Denklem 11 ve 12 kullanılarak hesaplanır. Denklem 9'da α yerine $\alpha/2$ yerleştirilip gerekli işlemler yapıldıktan sonra, ikinci gruba ait örneklem büyüklüğü

$$f(n_2) = \left(\frac{1}{\kappa n_2 - 3} + \frac{1}{n_2 - 3}\right) - \frac{(z_1 - z_2)^2}{\left(z_{\alpha/2} + z_\beta\right)^2} = 0 \qquad (90)$$

denkleminden R programındaki `uniroot()` kök bulma algoritması kullanılarak bulunur. $\kappa = n_1/n_2$ olduğundan birinci grup için örneklem büyüklüğü $n_1 = \kappa n_2$ eşitliğinden bulunur.

Örnek: Bir önceki örnek tekrardan ele alınsın fakat bu sefer, bireyci toplumlarda işbirlikçi öğrenme ortamı ve okul aidiyeti arasında $r_1 = 0.23$ büyüklüğünde pozitif bir ilişki olabileceği gibi $r_1 = 0.27$ büyüklüğünde pozitif bir ilişki de olabilir. Bir başka deyişle, r_1 ve r_2 arasindaki 0.02 birimlik fark negatif ya da pozitif olabilir. Her bir gruptan 50'şer kişilik bir örneklem ile istatistiksel güç ya da %80 güç için örneklem büyüklüğü `alternative="not equal"` olacak şekilde tanımlanarak

```
pwrss.z.2corrs(r1 = .23, r2 = 0.25,
               n2 = 50, alternative = "not equal")
# Difference between two correlations
# (independent samples z test)
# H0: r1 = r2
# HA: r1 != r2
# ------------------------------
#  Statistical power = 0.051
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = -0.103
# Type I error rate = 0.05
# Type II error rate = 0.949

pwrss.z.2corrs(r1 = .23, r2 = 0.25,
               power = 0.8, alternative = "not equal")
# Difference between two correlations
# (independent samples z test)
# H0: r1 = r2
# HA: r1 != r2
# ------------------------------
#  Statistical power = 0.8
#  n1 = 34854
#  n2 = 34854
# ------------------------------
```

```
# Alternative = "not equal"
# Non-centrality parameter = -2.802
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanır ve sonuçlar elde edilir. Buna göre, bireyci toplumlardaki korelasyonun kolektivist toplumlardaki korelasyondan farklı olduğunu ve 0.02 birimlik bir farkın pratikte anlamlı olduğunu iddia eden bir araştırmacı, her bir grupta 50'şer kişi ile hipotez testini %5.1 güç oranı ile test edebilir. Hipotez testini %80 güç oranı ile gerçekleştirmek için her bir gruptan en az 34854 katılımcıya gerek vardır.

Çoklu Doğrusal Regresyonda R^2 (veya ΔR^2)

İlgili tüm değişkenler regresyon modeline eklenip yordanan değişkendeki açıklanan varyans oranı araştırılabileceği gibi, değişkenler set halinde eklenip (hiyerarşik regresyon analizi) açıklanan varyanstaki değişim de araştırılabilir. Yokluk (H_0) ve alternatif (H_A) hipotezleri

H_0: $\Delta R^2 = 0$

H_A: $\Delta R^2 > 0$

şeklinde gösterilir. Açıklanan varyansın açıklanamayan varyansa oranı

$$f^2 = \frac{\Delta R^2}{1 - \Delta R^2} \tag{91}$$

şeklinde ifade edilir (Cohen f^2; Cohen, 1988). Gözlem sayısı n olan bir örneklemde, toplamda k adet yordayıcı değişkenden m âdetinin incelenecek değişken seti olduğu farz edilsin. Test istatistiği

$$\lambda = f^2 n \tag{92}$$

olarak ifade edilir. Yokluk hipotezi doğru olduğunda bu istatistik merkezi 0, pay için serbestlik derecesi $u = m$ ve payda için serbestlik derecesi $v = n - k - 1$ olan F dağılımını izler. Alternatif hipotez doğru olduğunda ise istatistik aynı serbestlik derecelerine sahip fakat merkezi λ olan F dağılımını izler. F dağılımının kümülatif dağılım fonksiyonu Φ_F ile, ters kümülatif dağılım fonksiyonu ise Φ_F^{-1} ile gösterilirse istatiksel gücü hesaplamak için

$$F_k = \Phi_F^{-1}(\alpha, u, v;\ 0) \tag{93}$$

$$1 - \beta = \Phi_F(F_k, u, v;\ \lambda) \tag{94}$$

denklemleri kullanılır. Örneklem büyüklüğünü hesaplamak için Denklem 94 yeniden düzenlenip aşağıdaki gibi yazılır:

$$f(n) = \Phi_F(F_k, u, v;\ \lambda) + \beta - 1 = 0 \qquad (95)$$

Daha sonra, R programındaki `uniroot()` kök bulma algoritması kullanılarak belirlenen güç oranına denk gelen örneklem büyüklüğü bulunur.

Örnek: Bir araştırmacı, COVID-19 döneminde, demografik özelliklerin (yaş ve cinsiyet) ötesinde, bazı savunma mekanizmalarının (arkadaşlarıyla konuşma, egzersiz, sosyal medya, kitap okuma, hobi, dini aktiviteler, alkol ve COVID-19 ile ilgili araştırmalar) psikosomatik belirtiler üzerindeki etkisini bulmak istemektedir. Daha önce yapılan benzer araştırmalar, regresyon modeline sadece yaş ve cinsiyet eklendiğinde ($m = 2$) psikosomatik belirtilerdeki varyansın sadece 0.01'nin açıklandığını, yaş ve cinsiyet ile birlikte sekiz savunma mekanizması eklendiğinde ise ($k = 8 + 2$) psikosomatik belirtilerdeki varyansın %24'nün açıklandığını bulmuşlardır (Otanga ve diğerleri., 2022). Modele sekiz değişken eklendiğinde açıklanan varyans değişimi $\Delta R^2 = 0.024 - 0.01 = 0.23$ olur. Burada açıklanan varyans değişimini tespit etmeye çalışan bir araştırmacı 50 kişi ile güç oranını ya da %80 güç için gerekli örneklem büyüklüğünü

```
pwrss.f.reg(k = 10, m = 8, n = 50, r2 = 0.23)
# R-squared change in hierarchical linear regression (F test)
# H0: r2 = 0
# HA: r2 > 0
# ------------------------------
#  Statistical power = 0.701
#  n = 50
# ------------------------------
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 39
# Non-centrality parameter = 14.935
# Type I error rate = 0.05
# Type II error rate = 0.299

pwrss.f.reg(k = 10, m = 8, power = 0.8, r2 = 0.23)
# R-squared change in hierarchical linear regression (F test)
# H0: r2 = 0
# HA: r2 > 0
# ------------------------------
#  Statistical power = 0.8
#  n = 59
# ------------------------------
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 25.385
# Non-centrality parameter = 10.868
# Type I error rate = 0.05
# Type II error rate = 0.2
```

kodlarını kullanarak hesaplayabilir. Sonuçlara göre, araştırmacı 50 katılımcı ile hipotez testini %70.1 güç oranı ile gerçekleştirilebilir. Öte yandan, hipotez testini %80 güç oranı ile gerçekleştirmek için en az 59 katılımcı gereklidir.

Şayet, araştırmacının odak noktası iki model arasındaki ΔR^2 (k = 10, m = 8) farkı değil de tek modelde R^2 değerinin (k = 10) 0'dan farkını test etmek ise örneklem büyüklüğü

```
pwrss.f.reg(k = 10, power = 0.8, r2 = 0.24)
# R-squared compared to 0 in linear regression (F test)
# H0: r2 = 0
# HA: r2 > 0
# ------------------------------
#  Statistical power = 0.8
#  n = 62
# ------------------------------
# Numerator degrees of freedom = 10
# Denominator degrees of freedom = 50.168
# Non-centrality parameter = 19.316
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplanabilirdi.

Varyans ve Kovaryans Analizi (ANOVA ve ANCOVA)

İki ya da fazla grup söz konusu olduğunda ve en az iki grup ortalaması arasında bir farkın olup olmadığı araştırılmak istendiğinde ANOVA (Varyans Analizi) testi kullanılır. Ortalamalar arası farklara bakılırken ön test gibi eş değişkenler de modele eklenebilir. Bu durumda ise ANCOVA (Kovaryans Analizi) testi kullanılır.

Bu hesaplamalarda Cohen f^2 etki büyüklüğü olarak kullanabileceği gibi kısmi η^2 (eta-kare) değeri de kullanılabilir. Raporlamalarda daha çok η^2 değeri kullanıldığı için buradaki hesaplamalarda da bu etki büyüklüğü türü kullanılacaktır. Yokluk (H_0) ve alternatif (H_A) hipotezlerini aşağıdaki gibi oluşturulur:

H_0: $\eta^2 = 0$ (ya da $f^2 = 0$)

H_A: $\eta^2 > 0$ (ya da $f^2 > 0$)

Yordanan değişkenin gruplar (ya da grup etkileşimleri) tarafından açıklanan varyansının kalan varyansa oranı

$$f^2 = \frac{\eta^2}{1 - \eta^2} \tag{96}$$

şeklinde hesaplanır (Cohen, 1988). Toplamda n tane katılımcıdan veri toplandığı, incelenecek faktörün k tane gruptan oluştuğu, grupların ortalamasının p tane eş değişken kullanılarak düzeltildiği farz edilsin. Bu durumda, test istatistiği

$$\lambda = f^2 n \tag{97}$$

olarak ifade edilir. Yokluk hipotezi doğru olduğunda bu istatistik merkezi 0, pay için serbestlik derecesi $u = k - 1$ ve payda için serbestlik derecesi $v = n - k - p$ olan F dağılımını izler. Alternatif hipotez doğru olduğunda ise istatistik aynı serbestlik derecelerine sahip fakat merkezi λ olan F dağılımını izler. Eş değişkenler modele eklenmediğinde, yani $p = 0$ olduğunda, grup karşılaştırmaları ANOVA modeli kullanılarak yapılır. Şayet η^2ya da f^2 düzeltilmiş ortalama farklarından elde edilmişlerse $p > 0$ olmalıdır. Bu durumda kullanılan model ANCOVA modelidir. ANOVA ya da ANCOVA modeli birden fazla faktörden oluşabilir (tek faktör, iki faktörlü veya üç faktörlü), bu faktörlerden her biri farklı sayıda gruplardan oluşabilir ve bu faktörler arası etkileşimler incelenmek istenebilir.

İstatiksel güç Denklem 93 ve 94 kullanılarak hesaplanır. Örneklem büyüklüğünü hesaplamak için ise R programındaki `uniroot()` kök bulma algoritması kullanılarak Denklem 95'teki şartı sağlayan n belirlenir.

Örnek: Aslan (2019) argümantasyona dayalı öğretim ve senaryo temelli öğrenme yöntemlerinin etkililiklerini bulmaya çalıştığı araştırmasında ANCOVA testini uygulamış ve gruplar (iki deney bir kontrol) için η^2 değerini 0.14 olarak bulmuştur. Toplamda 50 katılımcı ile güç oranı veya %80 güç için gerekli toplam örneklem büyüklüğü

```
pwrss.f.ancova(eta2 = 0.14, n = 50,
               n.way = 1, n.levels = 3, n.covariates = 1)
# One-way Analysis of Covariance (ANCOVA)
#  H0: 'eta2' or 'f2' = 0
#  HA: 'eta2' or 'f2' > 0
# ------------------------------------
# Factor A: 3 levels
# ------------------------------------
# Given eta2 = 0.14 or f2 = 0.163
#  Statistical power = 0.695
# ------------------------------------
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 46
# Non-centrality parameter = 8.14

pwrss.f.ancova(eta2 = 0.14, power = 0.8,
               n.way = 1, n.levels = 3, n.covariates = 1)
# One-way Analysis of Covariance (ANCOVA)
#  H0: 'eta2' or 'f2' = 0
```

```
#  HA: 'eta2' or 'f2' > 0
# ----------------------------------
# Factor A: 3 levels
# ----------------------------------
# Given eta2 = 0.14 or f2 = 0.163
#  Total n = 63
# ----------------------------------
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 58.33
# Non-centrality parameter = 10.15
```

şeklinde hesaplanır ve sonuçlar elde edilir. Sadece ön testin eş değişken olarak modele eklendiği düşünülerek `n.covariates=1` olarak tanımlanmıştır. `n.levels` argümanı faktörün kaç tane düzeyden (gruptan) oluştuğunu tanımlamak için kullanılır; faktör iki deney bir kontrol gruplardan oluştuğu için 3 olarak tanımlanmıştır. Bunlara ek olarak, iki faktörlü ANOVA ya da ANCOVA analizleri için örnek tanımlamalar şu şekilde yapılabilir: `n.levels=c(3,2)`. Bu tanımlama, ANOVA ya da ANCOVA modelinde iki faktörün kullanılacağını, birinci faktörün 3 düzeyden (gruptan) ve ikinci faktörün iki düzeyden (gruptan) oluştuğunu ifade etmektedir.

Tekrarlı Ölçümler Varyans Analizi (*Repeated Measures* ANOVA)

İki ya da daha fazla gruptan iki ya da daha fazla zaman noktasında tekrarlı ölçüm alındığı düşünülsün. Zaman etkisi kontrol edildiğinde en az iki grup ortalaması arasında bir fark olduğu, grup etkisi kontrol edildiğinde en az iki zaman noktası arasında bir fark olduğu ya da grup ve zaman etkileşimi olduğu düşünüldüğünde tekrarlı ölçümler ANOVA kullanılır. Cohen f^2 etki büyüklüğü olarak kullanabileceği gibi kısmi η^2 değeri de kullanılabilir. Yokluk (H_0) ve alternatif (H_A) hipotezleri aşağıdaki gibi oluşturulur:

H_0: $\eta^2 = 0$ (ya da $f^2 = 0$)

H_A: $\eta^2 > 0$ (ya da $f^2 > 0$)

Toplamda n tane katılımcıdan veri toplandığı, incelenecek faktörün k tane gruptan oluştuğu ve m tane ölçüm yapıldığı farz edilsin. Bu durumda, gruplar arası etkinin test istatistiği

$$\lambda = f^2 \left(\frac{m}{1 + (m-1)\rho} \right) n\epsilon \tag{98}$$

şeklinde ifade edilir. Zaman etkisi veya grup x zaman etkileşiminin test istatistiği ise

$$\lambda = f^2 \left(\frac{m}{1 - \rho} \right) n\epsilon \tag{99}$$

olarak ifade edilir. ϵ küresellik düzeltme faktörüdür ve $1/(m-1)$ ve 1 arasında değerler alır. Grup etkileri için; yokluk hipotezi doğru olduğunda bu istatistik merkezi 0, pay için serbestlik derecesi $u =$

$k-1$ ve payda için serbestlik derecesi $v=n-k$ olan F dağılımını izler; alternatif hipotez doğru olduğunda ise istatistik aynı serbestlik derecelerine sahip fakat merkezi λ (Denklem 98) olan F dağılımını izler.

Zaman etkisi için; yokluk hipotezi doğru olduğunda bu istatistik merkezi 0, pay için serbestlik derecesi $u=(m-1)\epsilon$ ve payda için serbestlik derecesi $v=(n-k)(m-1)\epsilon$ olan F dağılımını izler; alternatif hipotez doğru olduğunda ise istatistik aynı serbestlik derecelerine sahip fakat merkezi λ (Denklem 99) olan F dağılımını izler.

Grup ve zaman etkileşimi için; yokluk hipotezi doğru olduğunda bu istatistik merkezi 0, pay için serbestlik derecesi $u=(k-1)(m-1)\epsilon$ ve payda için serbestlik derecesi $v=(n-k)(m-1)\epsilon$ olan F dağılımını izler; alternatif hipotez doğru olduğunda ise istatistik aynı serbestlik derecelerine sahip fakat merkezi λ (Denklem 99) olan F dağılımını izler.

İstatiksel güç Denklem 93 ve 94 kullanılarak hesaplanabilir. Örneklem büyüklüğünü hesaplamak için ise R programındaki `uniroot()` kök bulma algoritması kullanılarak Denklem 95'teki şartı sağlayan n belirlenir.

Örnek: Kartal ve diğerleri. (2016) okul öncesi ve birinci sınıfların fonolojik farkındalıklarını arttırmayı hedefleyen iki farklı deney grubu (sınıf içi oyun veya bilgisayar destekli) ve bir de kontrol grubu oluşturup üç zaman noktasında ölçümler gerçekleştirmiştir (ön-test, son-test ve izleme testi). Kartal ve diğerleri. (2016) okul öncesi grubunda 53 katılımcıyı incelemiş, zaman etkisini η_Z^2 = 0.56, grup etkisini η_G^2 = 0.47 ve zaman x grup etkileşimini ise η_{ZxG}^2 = 0.10 bulmuştur. Öncelikli olarak, gruplar arası farkları tespit etmek isteyen bir araştırmacı, 53 katılımcı ile testin istatistiksel gücünü ve %80 istatistiksel güç için örneklem büyüklüğünü

```
##--------------- grup etkisi ----------------##
pwrss.f.rmanova(eta2 = 0.47, corr.rm = .5,
                n.levels  = 3, n.rm = 3,
                alpha = 0.05, n = 53, type = "between")
# One-way repeated measures analysis of variance (F test)
# H0: eta2 = 0 (or f2 = 0)
# HA: eta2 > 0 (or f2 > 0)
# ------------------------------
# Number of levels (groups) = 3
# Number of measurement time points = 3
# ------------------------------
#  Statistical power = 1
#  Total n = 53
# ------------------------------
# Type of the effect = "between"
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 50
```

```
# Non-centrality parameter = 70.5
# Type I error rate = 0.05
# Type II error rate = 0

pwrss.f.rmanova(eta2 = 0.47, corr.rm = .5,
                n.levels  = 3, n.rm = 3,
                alpha = 0.05, power = 0.8, type = "between")
# One-way repeated measures analysis of variance (F test)
# H0: eta2 = 0 (or f2 = 0)
# HA: eta2 > 0 (or f2 > 0)
# ------------------------------
# Number of levels (groups) = 3
# Number of measurement time points = 3
# ------------------------------
#  Statistical power = 0.8
#  Total n = 11
# ------------------------------
# Type of the effect = "between"
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 7.871
# Non-centrality parameter = 14.46
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplayabilir. Sonuçlardan görüldüğü üzere, araştırmacı gruplar arası farkları tespit edebilmek için 53 katılımcıdan veri toplayacak olursa hipotez testinin sahip olacağı güç oranı %100 iken, %80 güç oranı ile hipotez testini gerçekleştirmek için sadece 11 katılımcıya ihtiyacı vardır.

Grup etkisinden bağımsız olarak, zamana bağlı bir farklılığı tespit etmeye çalışan bir araştırmacı 53 katılımcı ile testin istatistiksel gücünü ve %80 istatistiksel güç için örneklem büyüklüğünü ise

```
##-------------- zaman etkisi ---------------##
pwrss.f.rmanova(eta2 = 0.56, corr.rm = .5,
                n.levels  = 3, n.rm = 3,
                alpha = 0.05, n = 53, type = "within")
# One-way repeated measures analysis of variance (F test)
# H0: eta2 = 0 (or f2 = 0)
# HA: eta2 > 0 (or f2 > 0)
# ------------------------------
# Number of levels (groups) = 3
# Number of measurement time points = 3
# ------------------------------
#  Statistical power = 1
```

```
#  Total n = 53
# ------------------------------
# Type of the effect = "within"
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 100
# Non-centrality parameter = 404.727
# Type I error rate = 0.05
# Type II error rate = 0

pwrss.f.rmanova(eta2 = 0.56, corr.rm = .5,
                n.levels  = 3, n.rm = 3,
                alpha = 0.05, power = 0.8, type = "within")
# One-way repeated measures analysis of variance (F test)
# H0: eta2 = 0 (or f2 = 0)
# HA: eta2 > 0 (or f2 > 0)
# ------------------------------
# Number of levels (groups) = 3
# Number of measurement time points = 3
# ------------------------------
#  Statistical power = 0.8
#  Total n = 5
# ------------------------------
# Type of the effect = "within"
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 2.826
# Non-centrality parameter = 33.699
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplayabilir. Sonuçlara göre, araştırmacı zamana bağlı farkları tespit edebilmek için 53 katılımcıdan veri toplayacak olursa hipotez testinin sahip olacağı güç oranı %100 iken, %80 güç oranı ile hipotez testini gerçekleştirmek için sadece 5 katılımcıya ihtiyacı vardır.

Son olarak, grup ve zaman etkileşimini tespit etmeye çalışan bir araştırmacı 53 katılımcı ile testin istatistiksel gücünü ve %80 istatistiksel güç için örneklem büyüklüğünü

```
##----- grup x zaman etkileşimi etkisi -------##
pwrss.f.rmanova(eta2 = 0.1, corr.rm = .5,
                n.levels  = 3, n.rm = 3,
                alpha = 0.05, n = 53, type = "interaction")
# One-way repeated measures analysis of variance (F test)
# H0: eta2 = 0 (or f2 = 0)
# HA: eta2 > 0 (or f2 > 0)
# ------------------------------
# Number of levels (groups) = 3
```

```
# Number of measurement time points = 3
# ------------------------------
#  Statistical power = 0.999
#  Total n = 53
# ------------------------------
# Type of the effect = "interaction"
# Numerator degrees of freedom = 4
# Denominator degrees of freedom = 100
# Non-centrality parameter = 35.333
# Type I error rate = 0.05
# Type II error rate = 0.001

pwrss.f.rmanova(eta2 = 0.10, corr.rm = .5,
                n.levels  = 3, n.rm = 3,
                alpha = 0.05, power = 0.8, type = "interaction")
# One-way repeated measures analysis of variance (F test)
# H0: eta2 = 0 (or f2 = 0)
# HA: eta2 > 0 (or f2 > 0)
# ------------------------------
# Number of levels (groups) = 3
# Number of measurement time points = 3
# ------------------------------
#  Statistical power = 0.8
#  Total n = 21
# ------------------------------
# Type of the effect="interaction"
# Numerator degrees of freedom = 4
# Denominator degrees of freedom = 34.973
# Non-centrality parameter = 13.658
# Type I error rate = 0.05
# Type II error rate = 0.2
```

şeklinde hesaplayabilir. Hesaplamalara göre, araştırmacı grup ve zaman etkileşimini tespit edebilmek için 53 katılımcıdan veri toplayacak olursa hipotez testinin sahip olacağı güç oranı %99.9 iken, %80 güç oranı ile hipotez testini gerçekleştirmek için sadece 21 katılımcıya ihtiyacı vardır.

Karmaşık Araştırma Desenlerinde Yeni Yaklaşımlar

Son zamanlarda güç analizi programlarında hızlı bir artış yaşanmaktadır. Özellikle, deneysel, yarı-deneysel, ve zayıf deneysel çalışma (basit ve çok düzeyli) alanlarında güç oranı ve örneklem büyüklüğünü hesaplamaları için formüller türetilmiş (Bloom, 1995, 2006, 2012; Bulus, 2022; Bulus ve Dong, 2021; Cattaneo, Titiunik, ve Vazquez-Bare, 2019; Dong ve diğerleri., 2021; Hedges ve Rhoads, 2010; Kelcey ve diğerleri., 2017a, 2017b; Konstantopoulos, 2008a, 2008b; Schochet, 2008, 2009; ve

diğerleri) ve bu formüller Excel dosyalarında (Dong ve Maynard, 2013), R paketlerinde (Bulus ve Dong, 2021; Bulus ve diğerleri., 2021; Cattaneo ve diğerleri., 2019), ve web uygulamalarında (bkz. Tablo 1) araştırmacıların kolay erişimine sunulmuştur.

Bunların yanı sıra, deneysel çalışmalarda sadece müdahale etkisi değil, müdahale etkisinde düzenleyici ve aracılık rolü olan değişkenler için de güç oranı ve örneklem büyüklüğü hesaplamaları yapılabilir. Bilindiği kadarıyla hesaplama programı olarak bu üçünü de kapsamlı bir şekilde uygulayan sadece PowerUp! Excel dosyaları (https://www.causalevaluation.org/power-analysis.html) ve PowerUpR paketi (Bulus ve diğerleri., 2021) bulunmaktadır.

Tablo 1. *Web Tabanlı Bazı Açık Erişim Güç Analizi Programları*

Açıklama	Link
Genel amaçlı hipotez testleri	https://pwrss.shinyapps.io/index/ https://pwrss.shinyapps.io/lang-en/ https://pwrss.shinyapps.io/lang-tr/
Genel amaçlı hipotez testleri	http://biostatapps.inonu.edu.tr/WSSPAS/
Genel amaçlı hipotez testleri	http://powerandsamplesize.com/
Genel amaçlı hipotez testleri	https://webpower.psychstat.org/wiki/models/index/
Çok düzeyli seçkisiz deneysel desenler	https://powerupr.shinyapps.io/index/
Çok düzeyli regresyon süreksizliği tasarımları	https://cosa.shinyapps.io/index/
Yapısal Eşitlik Modellemesi	https://yilinandrewang.shinyapps.io/pwrSEM/
Aracılık Analizi	https://davidakenny.shinyapps.io/MedPower/

Wang ve Rhemtulla (2021) yapısal eşitlik modellemesinde (YEM) tahmin edilen parametrelerin güç oranını hesaplamak için Monte-Carlo (MC) simülasyonuna dayalı R tabanlı pwrSEM uygulamasını geliştirmişlerdir. Yine *Mplus* (Muthén ve Muthén, 1998-2015) programındaki MONTECARLO komutu ile herhangi bir modelde istenilen herhangi bir parametrenin güç oranı hesaplanabilir. Daha basit aracılık modelleri için David A. Kenny'nin web uygulaması da kullanılabilir (bkz. Tablo 1). Düzenleyici değişkenlerin ilgilenildiği basit regresyon modellerinde ise R programında InteractionPoweR paketi kullanışlı olabilir (Baranger ve diğerleri., 2022).

Ahi Evran Üniversitesi
Kırşehir Eğitim Fakültesi Dergisi
ISSN: 2147 - 1037
http://kefad.ahievran.edu.tr

ENGLISH VERSION

Introduction

The population is the community encompassing the entire set of subjects that interest the researcher. Since resources are limited, reaching all the units that make up the population is often almost impossible. Therefore, data from a subset that contains and represents characteristics of the sample is collected and analyzed. This representative subset smaller than the population, which is more economical and manageable, is called a *sample*. One of the fundamental problems of statistics is the question of what the minimum sample size should be to make reliable inferences. Before examining this problem in depth, it is necessary to explain some basic concepts in statistics.

The value of any feature (variable or relationship between variables) in the population is called a *parameter*, and the value obtained from the sample is called a *statistic*. The statistic obtained from the sample is an estimate of the population parameter. Since not all units in the population are selected, we cannot claim that the statistic and parameter will be the same; nonetheless, these two values are expected to be close. Moreover, when a new sample is selected, the statistic may be different from the earlier one and may vary from one sample to another. These deviations from the population parameter are due to sampling error and are expressed as *standard errors* of statistics when statistical models are correctly specified.

In scientific studies, the standard error of a statistic is reported together with the statistic obtained from the sample. If information about the standard error is available, a sample size that will keep the standard error at a reasonable level may be determined prior to the study. It is impossible to obtain a sample of an infinite number of units. A sample consisting of a single unit or in which a single observation was made is also unacceptable. An unduly small sample makes it difficult to detect important effects (or differences) in practice. Therefore, the resources are wasted, and the participants take unnecessary risks. In an unduly large sample, too many resources are used to find insignificant effects in practice, and too many participants may take unnecessary risks. For ethical and economic reasons, it is necessary to determine the sample size to keep the standard error reasonable. The minimum required sample size for research is determined via *statistical power analysis*.

It is crucial to perform statistical power analysis in line with international standards for reporting scientific studies (e.g., *What Works Clearinghouse, Strengthening the Reporting of Observational Studies in Epidemiology, Consolidated Standards of Reporting Trials*). Although there are numerous sources on power analysis in the literature (e.g., Aberson, 2019; Cohen, 1988; Hedberg, 2017; Liu, 2013; Myors et al., 2023; Zhang and Yuan, 2018), this issue is not given enough importance in Türkiye, especially in the field of social sciences and humanities. Examining a representative sample of experimental studies reported in the educational and psychological sciences in Türkiye between 2010 and 2020, Bulus and Koyuncu (2021) found that none of the 155 experimental studies included power analysis calculations to determine the sample size. Similarly, Şevgin and Çetin (2017) randomly selected three of the journals in the field of educational sciences in Türkiye. They examined 25 quantitative studies published in these journals between 2014 and 2016, and as a result, none of them performed a power analysis.

Although there have been recent initiatives in Türkiye to provide open-access power analysis calculation tools, especially in biostatistics (e.g., Arslan et al., 2018), these endeavors are not reflected in education and behavioral sciences. Therefore, this study aims to explain the theoretical foundations and computational approach of statistical power analysis in light of commonly used hypothesis tests and provide practical examples from education and behavioral sciences.

Parameters to Consider in Power Analysis

We need objective criteria to determine the sample size to keep the standard error at a reasonable level by power analysis. These can be listed as type I error, type II error, the direction of hypothesis testing, a minimum effect that is important in a practical sense, a cutoff below which any effect would be ignorable in a practical sense (margin), and type of hypothesis test.

Type I and Type II Error

Since the population parameter is not known in reality, the alternative hypothesis (H_A) may be true, as well as the null hypothesis (H_0). There may be inference errors depending on which one is true. The null hypothesis may be true in the population but rejected in the sample, or the null hypothesis may not be rejected in the sample while the alternative hypothesis is true in the population. Inferring the presence of an effect in the sample that is not present in the population (H_0 is true, H_A is false), i.e., falsely rejecting the null hypothesis, is called a *type I error* (α). In cases where there are no multiple comparisons and multiple outcomes, it is usually specified as $\alpha = 0.05$. This value means, for example, that when 100 samples are drawn hypothetically, we can tolerate committing type I errors in up to 5 of them.

Inferring that there is no effect in the sample, whereas the effect exists in the population (H_0 is false, H_A is true), i.e., failing to reject the null hypothesis falsely, is called a *type II error* (β). It is often specified as $\beta = 0.20$. This value means that out of 100 hypothetical samples, we can tolerate committing type II errors in up to 20 of them. *Statistical power* (1 - β) is the probability of inferring that an effect that

exists in the population also exists in the sample (H_0 is false, H_A is true). For example, when statistical power is 0.80, it means that out of 100 hypothetical samples, we conclude that the effect exists in at least 80 of them.

Direction of Hypothesis Testing

When performing hypothesis tests, the ratio of an estimate minus the reference value (often, the value of the null) to the estimate's standard error is called a *test statistic* (e.g., observed or calculated *z* or *t* value). The test statistic is compared to a cutoff value (e.g., critical *z* or *t* value) corresponding to a given distribution's type I error rate (e.g., standard normal distribution or *t* distribution). The decision to reject or fail to reject the null hypothesis relies on comparing the test statistics and the critical value.

The type I error rate is also a function of the directionality of the hypothesis test. In unidirectional hypothesis testing, one believes that the estimate obtained from the sample is either less or greater than the reference value suggested by the null hypothesis. In bidirectional hypothesis testing, one believes that the estimate obtained from the sample is different from the reference value indicated by the null hypothesis (could be less as well as greater). For example, presume that the type I error is set at 0.05. For the unidirectional hypothesis test, the critical value is determined from a probability of 0.05. In this case, we believe that the critical value is on one side of the null distribution (one-tailed), and the probability of observing test statistics on the null distribution equal or greater (or less) to the critical value is 0.05. For the bidirectional hypothesis test, the critical value is determined from a probability of 0.025. In this case, we believe that the critical value is on two sides of the null distribution (two-tailed), and the probability of observing test statistics on the null distribution equal or less than the critical value on the left is 0.025 and equal or greater than the critical value on the right is 0.025.

While the value of 0 is often assigned to the null hypothesis (and is the default on most software programs), there are cases where the reference value can be set to a small negligible value (a.k.a. margin) that can be considered null in practical terms. In this line, *non-inferiority* (equal or greater in a practical sense), *superority* (greater in a practical sense), and *equivalence* (equal in a practical sense) hypothesis tests are primarily used in medical and pharmaceutical research but can also be helpful in educational, behavioral, and social sciences. These types of tests influence the test statistics. Numerous sources are available containing details and interpretations of such tests (e.g., Bokai, Hongyue, Xin, and Changyong, 2017; CPMP, 1998, 2001; Serdar, Cihan, Yücel and Serdar, 2021)

Smallest Meaningful Effect

We also need the smallest meaningful effect in practice to calculate the statistical power or sample size. One can decide on the smallest meaningful effect based on results from existing studies, experts, and reports. For example, the smallest meaningful effect can be related to the least but significant improvement in the symptoms of patients with depression or the least but meaningful improvement in the achievement of students.

The practice of using effect sizes reported in previous studies in power calculations has been criticized by some researchers (Bulus and Koyuncu, 2021; Gelman, 2019). An effect reported in previous studies may not be the smallest meaningful effect. Nonetheless, the effect size reported in previous studies can be used when investigating whether a new program is at least as effective as the older program or when exploring the reproducibility of results in earlier studies.

Type of Hypothesis Test

Although the rationale to calculate statistical power or minimum required sample size is similar regardless of the type of hypothesis test (*t*, *z*, *F*, etc.), there are slight differences between them. We need type I error, degrees of freedom, and direction or type of hypothesis test to determine the critical *t* value (and to calculate statistical power). Iterative root-finding algorithms are used to calculate the sample size because the critical *t* value depends on the degrees of freedom, and the degrees of freedom depend on the sample size. On the contrary, we need only the type I error and direction or type of hypothesis test to determine the critical *z* value. Since the critical *z* value is not affected by the sample size, the sample size is calculated directly by formula without the need for iterative root-finding algorithms. We need type I error and degrees of freedom (for numerator and denominator) to find the critical *F* value (and to calculate statistical power). However, iterative root-finding algorithms are used to determine the sample size because the degrees of freedom for the numerator and denominator depend on the number of groups or measurements and the sample size.

Statistical Power Analysis in R

Although there are many great programs for statistical power and sample size calculations (e.g., pwr R package, Champley et al., 2020; G*Power, Erdfelder et al., 1996), the availability of web applications in multiple languages makes pwrss R package an attractive option (Bulus, 2023). The following code group can be used to install and activate the package in the R environment.

```
# Install
install.packages("pwrss")
# Activate
library(pwrss)
```

The following sections will explain the formulas and equations required for power analysis and then show how to perform the calculations using practical examples.

Comparing a Single Proportion to a Constant

Comparison of a proportion ($\hat{p}$) obtained from a sample with a fixed proportion (p_0) can be performed by the *z* test. The standard error can be formulated as $SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$ where n is the sample size, however, the problem is that the standard error depends on the estimated proportion. When the estimated proportion is $\hat{p} \cong 0.50$, the standard error is relatively large. The standard error becomes smaller towards extremes. Cohen (1988) transformed proportions using the inverse of the sine

function to overcome this situation and proposed to perform statistical operations on these transformed values. Inverse sine function transformation of an estimate $\hat{p}$ is

$$\phi_{\hat{p}} = 2\arcsin(\hat{p}) \tag{1}$$

The standard error of the transformed value is $SE(\hat{\phi}) = \sqrt{1/n}$. Then, the test statistic is calculated as

$$z = \frac{\phi_{\hat{p}}}{\sqrt{1/n}} \tag{2}$$

When the test statistic is known, statistical power can be easily calculated.

One-way hypothesis testing: Considering that a proportion in the population (p) is smaller or greater than a fixed proportion (p_0), the null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $p \geq p_0$ (or $p \leq p_0$)

H_A: $p < p_0$ (or $p > p_0$)

The inverse sine function transformations and test statistics are calculated as in Equations 3-6. Test statistics is calculated as

$$\phi_{\hat{p}} = 2\arcsin(\hat{p}) \tag{3}$$

$$\phi_{p_0} = 2\arcsin(p_0) \tag{4}$$

$$\hat{h} = \phi_{\hat{p}} - \phi_{p_0} \tag{5}$$

$$z = \frac{\hat{h}}{SH(\hat{h})} = \frac{\phi_{\hat{p}} - \phi_{p_0}}{\sqrt{1/n}} \tag{6}$$

The critical value (z_k) can be obtained as in Equation 7 by defining the type I error rate (α) in the inverse cumulative density function of the standard normal distribution (Φ_z^{-1}). The statistical power ($1-\beta$) is obtained as in Equation 8 by defining z and z_k in the cumulative density function of the standard normal distribution (Φ_z) as

$$z_k = \Phi_z^{-1}(\alpha; 0) \tag{7}$$

$$1 - \beta = 1 - \Phi_z(z_k; z) \tag{8}$$

To calculate the minimum required sample size, an estimated test statistic corresponding to the desired type I error (α) and type II error (β) rates is calculated using Equation 9 ($z = z_\alpha + z_\beta$). Then, by rearranging Equation 6 we can obtain Equation 10 as

$$z_\alpha + z_\beta = \Phi_z^{-1}(\alpha; 0) + \Phi_z^{-1}(\beta; 0) \tag{9}$$

$$n = \frac{(z_\alpha + z_\beta)^2}{(\phi_{\hat{p}} - \phi_{p_0})^2} \quad (10)$$

Example: A researcher wonders if the proportion of children with learning difficulties in primary schools in a particular province is higher than that of the population, including all primary school students in the country. To do this, they are planning to randomly select 50 students from a list of all primary school pupils in the province. They will then calculate the proportion of children with learning difficulties ($\hat{p}$) and try to find out if this proportion is higher than the proportion in the population (p_0). The researcher will perform a one-way hypothesis test because they are convinced that his estimated proportion is not likely to be lower than the reference value in the population.

It is known that the proportion of children with learning difficulties in the population is around 0.06 (MEB, 2021). One of the parameters that the researcher should estimate before proceeding with the power analysis is the expected or minimum relevant proportion, and the other is how much type I error should be. The type I error rate is commonly taken as 0.05. Assuming that the expected or minimum relevant proportion is 0.10, with 50 participants, the statistical power is calculated as

```
pwrss.z.prop(p = 0.1, p0 = 0.06, alpha = 0.05,
             n = 50, alternative = "greater")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p = p0
# HA: p > p0
# ------------------------------
#  Statistical power = 0.276
#  n = 50
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 1.051
# Type I error rate = 0.05
# Type II error rate = 0.724
```

The power rate with 50 students is approximately 27.6%. A power rate of at least around 80% is widely accepted in the social and behavioral sciences. However, suppose severe financial, time, and personnel constraints or participants belong to a difficult-to-reach population. In that case, keeping the power rate above at least 50% may be permissible, provided methodological quality is not compromised. In this case, even if the study's results alone are unreliable, they may provide meaningful added value in combination with other studies when used in a meta-analysis later.

After all, to obtain a statistical power of 80%, more than 50 participants are needed. The minimum required sample size to meet these requirements is calculated as

```
pwrss.z.prop(p = 0.1, p0 = 0.06, alpha = 0.05,
             power = 0.8, alternative = "greater")
# Approach: Arcsine transformation
# One proportion compared to a constant (one sample z test)
# H0: p = p0
# HA: p > p0
# ------------------------------
#  Statistical power = 0.8
#  n = 281
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

The minimum required sample size is 281. If one firmly believes that the estimated proportion is expected to be smaller than the constant, they need to specify `alternative="less"`.

Two-way hypothesis testing: In two-way hypothesis testing, one believes that the expected or minimum relevant proportion (p) is different than the proportion in the population (p_0). The null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $p = p_0$

H_A: $p \neq p_0$

The test statistic is the same as in Equation 6. However, to consider the bidirectional nature of the hypothesis test, one should use $\alpha/2$ instead of α. Moreover, bidirectional hypothesis testing differs from one-way hypothesis testing in calculating statistical power. The critical value (z_k) and the statistical power ($1 - \beta$) are calculated as in Equations 11 and 12.

$$z_k = \Phi_z^{-1}(\alpha/2; 0) \tag{11}$$

$$1 - \beta = 1 - \Phi_z(z_k; z) + \Phi_z(-z_k; z) \tag{12}$$

The minimum required sample size is calculated as in Equations 9 and 10, but the type I error rate should be specified as $\alpha/2$ in the equations.

Example: Assume that a researcher is interested in whether the proportion of children with learning disabilities in the province differs from that of the whole population. Assuming, as in the previous example, that the estimated or minimum relevant proportion is 0.10, the statistical power with 50 participants or the minimum required sample size for a power rate of 80% is calculated by specifying `alternative="not equal"` as

```
pwrss.z.prop(p = 0.1, p0 = 0.06, alpha = 0.05, n = 50,
```

```
             alternative = "not equal")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p = p0
# HA: p != p0
# ------------------------------
#  Statistical power = 0.183
#  n = 50
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 1.051
# Type I error rate = 0.05
# Type II error rate = 0.817

pwrss.z.prop(p = 0.1, p0 = 0.06, alpha = 0.05, power = 0.8,
             alternative = "not equal")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p = p0
# HA: p != p0
# ------------------------------
#  Statistical power = 0.8
#  n = 356
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 2.802
# Type I error rate = 0.05
# Type II error rate = 0.2
```

In the one-way and two-way hypothesis tests described so far, the null hypothesis is rejected if the $\hat{p} - p_0$ difference is less than, greater than, or different from 0. Even if this difference is as small as 0.001, the null hypothesis is rejected as long as it is statistically significant. However, a minimal difference that is statistically significant may not make sense in practice. From this point of view, one can define a minimum value different from 0 called *margin*, and the hypothesis tests can be carried out accordingly. The margin is usually denoted by the symbol δ in the literature. Following this logic, there are also different types of one-way hypothesis testing, which are primarily used in medical or pharmaceutical research. The following paragraphs will describe one-way hypothesis tests of non-inferiority, superiority, and two one-way hypothesis tests of equivalence.

Non-inferiority or superiority hypothesis testing: For the non-inferiority test, the margin is usually negative when higher proportions express a positive phenomenon (e.g., the proportion of gifted students), and the null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $p - p_0 \leq \delta$

H_A: $p - p_0 > \delta$

When higher proportions indicate a negative phenomenon (e.g., the proportion of students with learning difficulties), the margin is usually positive, and the null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $p - p_0 \geq \delta$

H_A: $p - p_0 < \delta$

For the superiority test, the margin is usually positive in cases where higher proportions indicate a positive phenomenon (e.g., the proportion of gifted students), and the null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $p - p_0 \leq \delta$

H_A: $p - p_0 > \delta$

When higher proportions express a negative phenomenon (e.g., the proportion of students with learning difficulties), the margin is usually negative, and the null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $p - p_0 \geq \delta$

H_A: $p - p_0 < \delta$

The test statistics for both non-inferiority and superiority test can be expressed as

$$\phi_{p_0+\delta} = 2\arcsin(p_0 + \delta) \tag{13}$$

$$\hat{h} = \phi_{\hat{p}} - \phi_{p_0+\delta} \tag{14}$$

$$z = \frac{\hat{h}}{SH(\hat{h})} = \frac{\phi_{\hat{p}} - \phi_{p_0+\delta}}{\sqrt{1/n}} \tag{15}$$

Statistical power is calculated using Equations 7 and 8. The minimum required sample size is calculated using Equation 9 as

$$n = \frac{\left(z_\alpha + z_\beta\right)^2}{\left(\phi_{\hat{p}} - \phi_{p_0+\delta}\right)^2} \tag{16}$$

Example: A researcher wants to investigate whether the proportion of gifted children in a province is less than the proportion in the population ratio ($p_0 = 0.03$) in a practical sense. The value obtained from the sample can be slightly lower than the reference value ($\delta = -0.005$). In this case, the researcher can calculate the statistical power with 50 students or the minimum required sample size for a power rate of 80% by defining `margin=-0.005` and `alternative="non-inferior"` as

```
pwrss.z.prop(p = 0.04, p0 = 0.03, margin = -0.005,  alpha = 0.05,
             n = 50, alternative = "non-inferior")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p - p0 <= margin
# HA: p - p0 > margin
# ------------------------------
#  Statistical power = 0.398
#  n = 50
# ------------------------------
# Alternative = "non-inferior"
# Non-centrality parameter = 1.387
# Type I error rate = 0.05
# Type II error rate = 0.602

pwrss.z.prop(p = 0.04, p0 = 0.03, margin = -0.005, alpha = 0.05,
             power = 0.8, alternative = "non-inferior")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p - p0 <= margin
# HA: p - p0 > margin
# ------------------------------
#  Statistical power = 0.8
#  n = 161
# ------------------------------
# Alternative = "non-inferior"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

Per calculations, the statistical power with 50 participants is 39.8%. On the other hand, at least 161 participants are needed to perform this test with 80% statistical power.

Presume that the researcher wants to investigate whether the proportion of gifted children in a province is practically higher than that of the general population. The value obtained from the sample may be slightly higher than the reference value ($\delta = 0.005$), but this may not be considered higher in

practical terms. In this case, the researcher can find the statistical power with 50 students or the minimum required sample size for a power rate of 80% by defining `margin=0.005` and `alternative="superior"` as

```
pwrss.z.prop(p = 0.04, p0 = 0.03, margin = 0.005,  alpha = 0.05,
             n = 50, alternative = "superior")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p - p0 <= margin
# HA: p - p0 > margin
# ------------------------------
#  Statistical power = 0.012
#  n = 50
# ------------------------------
# Alternative = "superior"
# Non-centrality parameter = -0.615
# Type I error rate = 0.05
# Type II error rate = 0.988

pwrss.z.prop(p = 0.04, p0 = 0.03, margin = 0.005, alpha = 0.05,
             power = 0.8, alternative = "superior")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: p - p0 <= margin
# HA: p - p0 > margin
# ------------------------------
#  Statistical power = 0.8
#  n = 818
# ------------------------------
# Alternative = "superior"
# Non-centrality parameter = -2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

The statistical power with 50 participants is quite low (1.2%). At least 818 participants are needed to perform this test with 80% statistical power.

Equivalence hypothesis testing: Equivalence hypothesis testing is performed by using two one-way hypothesis tests. In this case, the null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $|p - p_0| \geq \delta$

H_A: $|p - p_0| < \delta$

Test statistics is calculated as

$$\hat{h} = |\phi_{\hat{p}} - \phi_{p_0+\delta}| \tag{17}$$

$$z = \frac{\hat{h}}{SH(\hat{h})} = \frac{|\phi_{\hat{p}} - \phi_{p_0+\delta}|}{\sqrt{1/n}} \tag{18}$$

The test statistic (z) and the critical value (z_k) are used to calculate the statistical power as

$$z_k = \Phi_z^{-1}(\alpha; 0) \tag{19}$$

$$1 - \beta = 2\big(1 - \Phi_z(z_k; z)\big) - 1 \tag{20}$$

Then, the minimum required sample size is

$$z_\alpha + z_{\beta/2} = \Phi_z^{-1}(\alpha; 0) + \Phi_z^{-1}(\beta/2; 0) \tag{21}$$

$$n = \frac{\left(z_\alpha + z_{\beta/2}\right)^2}{\left(|\phi_{\hat{p}} - \phi_{p_0+\delta}|\right)^2} \tag{22}$$

Example: Assume a researcher is investigating whether the proportion of gifted children in a province is practically equal to the proportion in the population. The value obtained from the sample may be slightly lower or higher than the population value ($\delta = 0.005$) but will not be considered lower or higher in practical terms. In this case, the statistical power with a sample size of 50 or the minimum required sample size for a power rate of 80% is calculated by defining `margin=0.005` and `alternative="equivalent"` as

```
pwrss.z.prop(p = 0.04, p0 = 0.03, margin = 0.005, alpha = 0.05,
             n = 50, alternative = "equivalent")
# Approach: Arcsine transformation
# Error: design is not feasible

pwrss.z.prop(p = 0.04, p0 = 0.03, margin = 0.005, alpha = 0.05,
             power = 0.80, alternative = "equivalent")
# Approach: Arcsine transformation
# One proportion compared to a constant
# (one sample z test)
# H0: |p - p0| >= margin
# HA: |p - p0| < margin
# ------------------------------
#  Statistical power = 0.8
#  n = 1132
# ------------------------------
# Alternative = “equivalent”
```

```
# Non-centrality parameter = -2.926
# Type I error rate = 0.05
# Type II error rate = 0.2
```

At least 1132 students are required to perform this test with a power rate of 80%. When calculating the statistical power in equivalence studies (the first code group), a warning may indicate that the design is not feasible. It means that it is impossible to calculate the power rate with 50 students.

Comparing Two Proportions

The previous section discussed testing a proportion obtained from a single sample against a constant value. Only a single proportion contributed to the standard error because the reference value being compared is a constant and does not change depending on the sample. On the other hand, one may want to compare two proportions from two different groups in a sample (e.g., $\hat{p}_1$ and $\hat{p}_2$) or from two different samples. In this case, since the proportions being compared are both estimated, both contribute to the standard error of the difference. The variances of the estimates are $\hat{p}_1(1-\hat{p}_1)$ and $\hat{p}_2(1-\hat{p}_2)$ respectively. Then, the standard error of difference can be stated as

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{23}$$

where n_1 and n_2 are sample sizes in the first and second groups. However, as mentioned in the previous section, the standard error of the difference depends on the estimates themselves. Proportions are transformed using the inverse sine function, and standard error is calculated as

$$\phi_{\hat{p}_1} = 2\arcsin(\hat{p}_1) \tag{24}$$

$$\phi_{\hat{p}_2} = 2\arcsin(\hat{p}_2) \tag{25}$$

$$\hat{h} = \phi_{\hat{p}_1} - \phi_{\hat{p}_2} \tag{26}$$

$$SE(\hat{h}) = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{27}$$

One-way hypothesis testing: One-way hypothesis test is conducted if one believes that the proportion of one group in the population (p_1) is less or greater than the proportion of the other group (p_2). The null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $p_1 \geq p_2$ (or $p_1 \leq p_2$)

H_A: $p_1 < p_2$ (or $p_1 > p_2$)

Test statistics is calculated as

$$z = \frac{\hat{h}}{SE(\hat{h})} = \frac{\phi_{\hat{p}_1} - \phi_{\hat{p}_2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{28}$$

Statistical power is calculated using Equations 7 and 8. After performing the operation in Equation 9, the minimum required sample size is calculated as

$$n_2 = \frac{\left(z_\alpha + z_\beta\right)^2}{\left(\phi_{\hat{p}_1} - \phi_{\hat{p}_2}\right)^2}\left(\frac{\kappa + 1}{\kappa}\right) \tag{29}$$

where κ is the ratio of the sample size of the first group to the sample size of the second group (n_1/n_2). The sample size for the first group can be found via $n_1 = n_2\kappa$.

Example: A researcher is trying to find out whether the proportion of boys with learning disabilities in primary schools in a particular province is higher than the proportion of girls with learning disabilities. They aim to collect data from 100 people, 50 from each group. Assume that the expected proportion of boys with learning disabilities (p_1) is 0.08, while for girls (p_2) it is 0.06. Statistical power with 50 students in each group and the minimum required sample size for a power rate of 80% is calculated as

```
pwrss.z.2props(p1 = 0.08, p2 = 0.06, alpha = 0.05,
               kappa = 1, n2 = 50, alternative = "greater")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: p1 = p2
# HA: p1 > p2
# ------------------------------
#  Statistical power = 0.105
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 0.393
# Type I error rate = 0.05
# Type II error rate = 0.895

pwrss.z.2props(p1 = 0.08, p2 = 0.06, alpha = 0.05,
               kappa = 1, power = 0.8, alternative = "greater")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: p1 = p2
# HA: p1 > p2
# ------------------------------
```

```
#  Statistical power = 0.8
#  n1 = 2003
#  n2 = 2003
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

Unlike the previous exercises, the `kappa=1` argument refers to the ratio of male and female participants (`n1/n2`). Collecting data from only 100 participants is insufficient because the statistical power is around 10.5%. If this test is to be performed with 80% power, data must be collected from 2003 participants from each group.

Two-way hypothesis testing: Two-way hypothesis testing can be performed if one believes that the proportion of one group in the population is thought to be different from the proportion of the other group. The null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $p_1 = p_2$

H_A: $p_1 \neq p_2$

The test statistic is calculated as in equation 28. Equations 11 and 12 are used to calculate the statistical power. After using $\alpha/2$ in Equation 9, the minimum required sample size for the second group is calculated as

$$n_2 = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2}{\left(\phi_{\hat{p}_1} - \phi_{\hat{p}_2}\right)^2}\left(\frac{\kappa + 1}{\kappa}\right) \tag{30}$$

Similarly, the sample size for the first group is calculated as $n_2 = n_1\kappa$.

Example: A researcher who is trying to find whether the proportion of boys with learning disabilities is not equal to the proportion of girls with learning disabilities will perform a two-way hypothesis test. Statistical power with 50 students in each group and minimum required sample size for a power rate of 80% can be found by specifying `alternative="not equal"` as

```
pwrss.z.2props(p1 = 0.08, p2 = 0.06, alpha = 0.05,
               kappa = 1, n2 = 50, alternative = "not equal")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: p1 = p2
# HA: p1 != p2
# ------------------------------
#  Statistical power = 0.068
```

```
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 0.393
# Type I error rate = 0.05
# Type II error rate = 0.932

pwrss.z.2props(p1 = 0.08, p2 = 0.06, alpha = 0.05,
               kappa = 1, power = 0.8, alternative = "not equal")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: p1 = p2
# HA: p1 != p2
# ------------------------------
#  Statistical power = 0.8
#  n1 = 2543
#  n2 = 2543
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 2.802
# Type I error rate = 0.05
# Type II error rate = 0.2
```

The statistical power of the hypothesis test with 50 participants in each group is 6.8%. At least 2543 participants are needed in each group to perform this test with 80% statistical power. The previously mentioned non-inferiority, superiority and equivalence hypothesis tests can also be established here.

Non-inferiority or superiority hypothesis testing: Non-inferiority or superiority hypothesis testing can be used when the proportion of one group in the population is believed to be equally good or better than the proportion of the other group while considering a margin (δ). For the non-inferiority test, the margin is usually negative when higher proportions indicate a positive phenomenon (e.g., the proportion of gifted students), and the null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $p_1 - p_2 \leq \delta$

H_A: $p_1 - p_2 > \delta$

In cases where higher proportions indicate a negative phenomenon (e.g., the proportion of students with learning difficulties), the margin is usually positive, and the null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $p_1 - p_2 \geq \delta$

$H_A: p_1 - p_2 < \delta$

For the superiority test, the margin is usually positive in cases where higher proportions indicate a positive phenomenon (e.g., the proportion of gifted students), and the null (H0) and alternative (HA) hypotheses are formed as follows:

$H_0: p_1 - p_2 \leq \delta$

$H_A: p_1 - p_2 > \delta$

In cases where higher proportions indicate a negative phenomenon (e.g., the proportion of students with learning difficulties), the margin is usually negative, and the null (H0) and alternative (HA) hypotheses are constructed as follows:

$H_0: p_1 - p_2 \geq \delta$

$H_A: p_1 - p_2 < \delta$

Test statistics calculated for both non-inferiority and superiority tests can be expressed as

$$z = \frac{\hat{h}}{SE(\hat{h})} = \frac{\phi_{\hat{p}_1} - \phi_{\hat{p}_2+\delta}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{31}$$

Equations 7 and 8 are used to calculate the statistical power. After the operation in Equation 9 is performed, the minimum required sample size for the second group is

$$n_2 = \frac{(z_\alpha + z_\beta)^2}{(\phi_{\hat{p}_1} - \phi_{\hat{p}_2+\delta})^2}\left(\frac{\kappa + 1}{\kappa}\right) \tag{32}$$

Similarly, the sample size for the first group is calculated as $n_1 = n_2\kappa$.

Example: A researcher wants to investigate whether the proportion of gifted boys in a province is practically lower than that of gifted girls. They will declare a practical difference if the difference between the two proportions is less than δ = -0.005. In this case, the statistical power of 50 participants from each group or the sample size for 80% statistical power is found by defining `margin=-0.005` and `alternative="non-inferior"` as

```
pwrss.z.2props(p1 = 0.04, p2 = 0.02, alpha = 0.05, margin = -0.005,
               kappa = 1, n2 = 50, alternative = "non-inferior")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: p1 - p2 <= margin
# HA: p1 - p2 > margin
# ------------------------------
```

```
#  Statistical power = 0.366
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "non-inferior"
# Non-centrality parameter = 1.302
# Type I error rate = 0.05
# Type II error rate = 0.634
```

```
pwrss.z.2props(p1 = 0.04, p2 = 0.02, alpha = 0.05,  margin = -0.005,
               kappa = 1, power = 0.8, alternative = "non-inferior")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: p1 - p2 <= margin
# HA: p1 - p2 margin
# ------------------------------
#  Statistical power = 0.8
#  n1 = 183
#  n2 = 183
# ------------------------------
# Alternative = "non-inferior"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

The statistical power for non-inferiority hypothesis testing with 50 participants in each group is 36.6%. At least 183 participants from each group are needed to achieve a power rate of 80%.

Presume that the researcher will investigate whether the proportion of gifted boys in a province is practically higher than that of gifted girls. They will reject the null when the difference between the two proportions is greater than $\delta = 0.01$. Statistical power with 50 participants from each group or the sample size for 80% power rate is found by defining `margin=0.01` and `alternative="superior"` as

```
pwrss.z.2props(p1 = 0.04, p2 = 0.02, alpha = 0.05, margin = 0.01,
               kappa = 1, n2 = 50, alternative = "superior")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: p1 - p2 <= margin
# HA: p1 - p2 > margin
# ------------------------------
#  Statistical power = 0.02
```

```
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "superior"
# Non-centrality parameter = -0.113
# Type I error rate = 0.05
# Type II error rate = 0.961

pwrss.z.2props(p1 = 0.04, p2 = 0.02, alpha = 0.05,  margin = 0.01,
               kappa = 1, power = 0.8, alternative = "superior")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: p1 - p2 <= margin
# HA: p1 - p2 > margin
# ------------------------------
#  Statistical power = 0.8
#  n1 = 1866
#  n2 = 1866
# ------------------------------
# Alternative = "superior"
# Non-centrality parameter = -2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

The power of the superiority hypothesis test with 50 participants in each group is only 2%. At least 1866 participants in each group are needed to perform this test with 80% statistical power. Note that the difference between non-inferiority and superiority is how margin (δ) is defined.

Equivalence hypothesis testing: Equivalence hypothesis testing is conducted when one believes that the proportion of one group in the population is equal to that of the other group while considering the margin (δ). The null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $|p_1 - p_2| \geq \delta$

H_A: $|p_1 - p_2| < \delta$

Test statistics is calculated as

$$z = \frac{\hat{h}}{SE(\hat{h})} = \frac{|\phi_{\hat{p}_1} - \phi_{\hat{p}_2+\delta}|}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{33}$$

Equations 19 and 20 are used to calculate the statistical power. After the operation in Equation 9 is performed, the minimum required sample size is

$$n_2 = \frac{(z_\alpha + z_\beta)^2}{\left(|\phi_{\hat{p}_1} - \phi_{\hat{p}_2+\delta}|\right)^2}\left(\frac{\kappa+1}{\kappa}\right) \quad (34)$$

Similarly, the sample size for the first group can be found as $n_1 = n_2\kappa$.

Example: A researcher wants to investigate whether the proportion of gifted boys in a province is practically equivalent to the proportion of gifted girls. The researcher can tolerate a difference between the two proportions as much as ∓ 0.01, yet they would claim equivalence. Again, the statistical power with 50 participants from each group or the sample size for 80% power is calculated via defining `margin=0.01` and `alternative="equivalent"` as

```
pwrss.z.2props(p1 = 0.04, p2 = 0.02, alpha = 0.05, margin = 0.01,
               n2 = 50, kappa = 1, alternative = "equivalent")
# Approach: Arcsine transformation
# Error: design is not feasible

pwrss.z.2props(p1 = 0.04, p2 = 0.02, alpha = 0.05,  margin = 0.01,
               kappa = 1, power = 0.8, alternative = "equivalent")
# Approach: Arcsine transformation
# Difference between two proportions
# (independent samples z test)
# H0: |p1 - p2| >= margin
# HA: |p1 - p2| < margin
# ------------------------------
#  Statistical power = 0.8
#  n1 = 2585
#  n2 = 2585
# ------------------------------
# Alternative = "equivalent"
# Non-centrality parameter = -2.926
# Type I error rate = 0.05
# Type II error rate = 0.2
```

The error in the first code chunk means it would not be sufficient to calculate the statistical power with 50 people in each group. In addition, 2585 participants from each group are required for the equivalence test with 80% statistical power.

Comparing a Single Mean to a Constant

Suppose there is a random variable X in the population with a mean μ and variance σ^2 and $x_1, x_2, x_3, \ldots, x_n$ values are observed in the sample. The mean and variance can be estimated via Equation 35 and 36

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{35}$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \tag{36}$$

The standard error is expressed as

$$SE(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{n}} \tag{37}$$

One-way hypothesis testing: If one believes that the mean in the population (μ) is less or greater than a constant (μ_0), the null (H_0) and alternative (H_A) hypotheses are established as follows:

H_0: $\mu \geq \mu_0$ (or $\mu \leq \mu_0$)

H_A: $\mu < \mu_0$ (or $\mu > \mu_0$)

The test statistics is calculated as

$$z = \frac{\hat{\mu} - \mu_0}{SE(\hat{\mu})} = \frac{\hat{\mu} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}} \tag{38}$$

Statistical power is calculated using Equations 7 and 8. The minimum required sample size is calculated as

$$n = \frac{\left(z_\alpha + z_\beta\right)^2 \hat{\sigma}^2}{(\hat{\mu} - \mu_0)^2} \tag{39}$$

Example: A researcher wants to find out whether the depression level of college students during COVID-19 is higher than 21 points (which can be considered a moderate level). The Beck Depression Inventory (BDE; Beck, Ward, Mendelson, Mock, and Erbaugh, 1961) consists of 21 items. Hisli (1989) stated that university students who scored 21 showed signs of moderate depression. Then, the reference value can be set to μ_0 = 21. Furthermore, Hisli (1989) found that the standard deviation of BDE scores of university students was 6.75. Assuming that a two-unit increase from the reference value (the moderate level), i.e., 23, is the minimum effect that is important for practice, the statistical power with 50 participants or the sample size for 80% power is calculated via defining `alternative="greater"` as

```
pwrss.z.mean(mu = 23, mu0 = 21, sd = 6.75,
             alpha = 0.05, n = 50, alternative = "greater")
# One mean compared to a constant
# (one sample z test)
# H0: mu = mu0
# HA: mu > mu0
# ------------------------------
#  Statistical power = 0.674
#  n = 50
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 2.095
# Type I error rate = 0.05
# Type II error rate = 0.326

pwrss.z.mean(mu = 23, mu0 = 21, sd = 6.75,
             alpha = 0.05, power = 0.8, alternative = "greater")
# One mean compared to a constant
# (one sample z test)
# H0: mu = mu0
# HA: mu > mu0
# ------------------------------
#  Statistical power = 0.8
#  n = 71
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

The power rate with 50 participants is 67.4%. In addition, if it is possible to collect data from more participants, at least 71 participants are needed to perform the hypothesis test with 80% statistical power.

Two-way hypothesis testing: A two-way hypothesis testing can be conducted when one believes that the population mean (μ) is not equal to a constant value (μ_0). The null (H_0) and alternative (H_A) hypotheses are established as follows:

H_0: $\mu = \mu_0$

H_A: $\mu \neq \mu_0$

The test statistic is calculated as in Equation 38. Equations 7 and 8 are used to calculate statistical power. The sample size is calculated as

$$n = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2 \hat{\sigma}^2}{(\hat{\mu} - \mu_0)^2} \tag{40}$$

Example: A researcher tries to prove that the level of depression in college participants during COVID-19 differs from the moderate level of 21 points. Assuming, as in the previous example, that the minimum significant difference is two units (i.e., the level of depression can be 19 units or 23 units), the statistical power with 50 participants or the sample size for 80% power is found by defining `alternative="not equal"` as

```
pwrss.z.mean(mu = 23, mu0 = 21, sd = 6.75,
             alpha = 0.05, n = 50, alternative = "not equal")
# One mean compared to a constant
# (one sample z test)
# H0: mu = mu0
# HA: mu != mu0
# ------------------------------
#  Statistical power = 0.554
#  n = 50
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 2.095
# Type I error rate = 0.05
# Type II error rate = 0.446

pwrss.z.mean(mu = 23, mu0 = 21, sd = 6.75,
             alpha = 0.05, power = 0.8, alternative = "not equal")
# One mean compared to a constant
# (one sample z test)
# H0: mu = mu0
# HA: mu != mu0
# ------------------------------
#  Statistical power = 0.8
#  n = 90
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 2.802
# Type I error rate = 0.05
# Type II error rate = 0.2
```

The power of the two-way hypothesis test with 50 participants is 55.4%. At least 90 participants are needed to perform the hypothesis test with 80% statistical power.

Non-inferiority or superiority hypothesis testing: Non-inferiority or superiority hypothesis testing is used when the mean in the population (μ) is thought to be not less than or greater than a constant value (μ_0) in the population while considering a margin (δ). For the non-inferiority test, where higher values of a variable indicate a positive phenomenon (e.g., achievement test score), the margin is usually negative, and the null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $\mu - \mu_0 \leq \delta$

H_A: $\mu - \mu_0 > \delta$

Where high values of the variable indicate a negative phenomenon (e.g., depression score), the margin is usually positive, and the null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $\mu - \mu_0 \geq \delta$

H_A: $\mu - \mu_0 < \delta$

For the superiority test, where high values of the variable indicate a positive phenomenon (e.g., achievement test score), the margin is usually positive, and the null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $\mu - \mu_0 \leq \delta$

H_A: $\mu - \mu_0 > \delta$

In cases where high values of the variable indicate a negative phenomenon (e.g., depression score), the margin is usually negative, and the null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $\mu - \mu_0 \geq \delta$

H_A: $\mu - \mu_0 < \delta$

Test statistics for both non-inferiority and superiority tests is calculated as

$$z = \frac{\hat{\mu} - \mu_0 - \delta}{SE(\hat{\mu})} = \frac{\hat{\mu} - \mu_0 - \delta}{\frac{\hat{\sigma}}{\sqrt{n}}} \tag{41}$$

Equations 7 and 8 are used for statistical power calculation. The minimum required sample size is calculated as

$$n = \frac{\left(z_\alpha + z_\beta\right)^2 \hat{\sigma}^2}{(\hat{\mu} - \mu_0 - \delta)^2} \tag{42}$$

Example: A researcher is trying to determine whether secondary school students' psychological resilience levels during COVID-19 are practically less than the value in a pre-COVID-19 article. They will reject the null hypothesis if the difference between the sample estimate and the constant value

exceeds -2. In the previous study, the average value of psychological resilience was determined as 49 units, and the standard deviation of psychological resilience scores was 7.59 (Arslan, 2015). Assuming that the expected mean is 51 units, statistical power with 50 participants or the minimum required sample size for a power rate of 80% is calculated by defining `margin=-2` ve `alternative="non-inferior"` as

```
pwrss.z.mean(mu = 51, mu0 = 49, sd = 7.59, margin = -2,
             alpha = 0.05, n = 50, alternative = "non-inferior")
# One mean compared to a constant
# (one sample z test)
# H0: mu - mu0 <= margin
# HA: mu - mu0 > margin
# ------------------------------
#  Statistical power = 0.981
#  n = 50
# ------------------------------
# Alternative = "non-inferior"
# Non-centrality parameter = 3.727
# Type I error rate = 0.05
# Type II error rate = 0.019

pwrss.z.mean(mu = 51, mu0 = 49, sd = 7.59, margin = -2,
             alpha = 0.05, power = 0.8, alternative = "non-inferior")
# One mean compared to a constant
# (one sample z test)
# H0: mu - mu0 <= margin
# HA: mu - mu0 > margin
# ------------------------------
#  Statistical power = 0.8
#  n = 23
# ------------------------------
# Alternative = "non-inferior"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

As a result, the power of the non-inferiority hypothesis test with 50 participants is 98.1%. At least 23 participants are required to perform the hypothesis test with 80% statistical power.

Equivalence hypothesis testing: Equivalence hypothesis testing can be considered when one believes that there is no difference between the population value (μ) and the fixed values (μ_0) while considering a margin (δ). The null (H_0) and alternative (H_A) hypotheses are formed as follows:

$$H_0: |\mu - \mu_0| \geq \delta$$

$H_A: |\mu - \mu_0| < \delta$

Test statistics is calculated as

$$z = \frac{|\hat{\mu} - \mu_0| - \delta}{SE(\hat{\mu})} = \frac{|\hat{\mu} - \mu_0| - \delta}{\frac{\hat{\sigma}}{\sqrt{n}}} \tag{43}$$

Equations 17 and 18 are used for statistical power calculation. The minimum required sample size is calculated via

$$n = \frac{(z_\alpha + z_{\beta/2})^2 \hat{\sigma}^2}{(|\hat{\mu} - \mu_0| - \delta)^2} \tag{44}$$

Example: A researcher is trying to determine whether the psychological resilience levels of middle school students equal the value in a pre-COVID-19 article. They will reject the null hypothesis if the difference between the sample estimate and the constant value is greater than -1 and less than 1. Statistical power with 50 participants or the minimum required sample size for a power rate of 80% is calculated by defining `margin=2` ve `alternative="equivalent"` as

```
pwrss.z.mean(mu = 51, mu0 = 49, sd = 7.59, margin = 1,
             alpha = 0.05, n = 50, alternative = "equivalent")
# Error: design is not feasible

pwrss.z.mean(mu = 51, mu0 = 49, sd = 7.59, margin = 1,
             alpha = 0.05, power = 0.8, alternative = "equivalent")
# One mean compared to a constant
# (one sample z test)
# H0: |mu - mu0| >= margin
# HA: |mu - mu0| < margin
# ------------------------------
#  Statistical power = 0.8
# n = 494
# ------------------------------
# Alternative = "equivalent"
# Non-centrality parameter = 2.926
# Type I error rate = 0.05
# Type II error rate = 0.2
```

The error given in the above output shows that it is impossible to determine the statistical power with a sample of 50 participants. At least 494 participants are required to perform the equivalence hypothesis test with 80% power.

Comparing Two Means

The independent samples *t*-test is used to compare the means of the two groups in cross-sectional data, and the dependent samples (matched pairs) *t*-test is used to compare the means of the same group at two different time points.

Independent samples t-test: Consider a random variable X_1 with a mean μ_1 and variance σ_1^2 for the first group, a random variable X_2 with a mean μ_2 and variance σ_2^2 for the second group. Assume that, in the sample, n_1 observations are realized for X_1 $(x_{11}, x_{12}, x_{13}, \dots, x_{1n_1})$ in the first group and n_2 observations are realized for X_1 $(x_{21}, x_{22}, x_{23}, \dots, x_{2n_2})$ in the second group. Means are estimated via

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i} \tag{45}$$

$$\hat{\mu}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i} \tag{46}$$

and variances are estimated via

$$\hat{\sigma_1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} (x_{1i} - \hat{\mu}_1)^2 \tag{47}$$

$$\hat{\sigma_2}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2} (x_{2i} - \hat{\mu}_2)^2 \tag{48}$$

Then, the standard error of the difference is calculated as

$$SE(\hat{\mu}_1 - \hat{\mu}_2) = \sqrt{\frac{\hat{\sigma_1}^2}{n_1} + \frac{\hat{\sigma_2}^2}{n_2}} \tag{49}$$

One-way hypothesis testing: One-way hypothesis testing is used if one believes that the mean of one group in the population (μ_1) is less or greater than the mean of the other group (μ_2). The null (H_0) and alternative (H_A) hypotheses are established as follows:

H_0: $\mu_1 \geq \mu_2$ (or $\mu_1 \leq \mu_2$)

H_A: $\mu_1 < \mu_2$ (or $\mu_1 > \mu_2$)

Test statistics is calculated as

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{SE(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^{\,2}}{n_1} + \frac{\hat{\sigma}_2^{\,2}}{n_2}}} \tag{50}$$

Then, statistical power is calculated as

$$v = n_1 + n_2 - 2 \tag{51}$$

$$t_k = \Phi_t^{-1}(\alpha, v; 0) \tag{52}$$

$$1 - \beta = 1 - \Phi_t(t_k, v; t) \tag{53}$$

where v refers to the degrees of freedom. The sample size is calculated as

$$t_\alpha + t_\beta = \Phi_t^{-1}(\alpha, v; 0) + \Phi_t^{-1}(\beta, v; 0) \tag{54}$$

$$n_2 = (t_\alpha + t_\beta)^2 \left(\frac{\frac{\hat{\sigma}_1^{\,2}}{\kappa} + \hat{\sigma}_2^{\,2}}{(\hat{\mu}_1 - \hat{\mu}_2)^2} \right) \tag{55}$$

Since $\kappa = n_1/n_2$, the sample size for the first group is obtained via $n_1 = n_2\kappa$. If statistical power or sample size calculations rely on the standardized mean difference such as Cohen's d, specify $\hat{\mu}_1 = d$, $\hat{\mu}_2 = 0$, $\hat{\sigma}_1^{\,2} = 1$ and $\hat{\sigma}_2^{\,2} = 1$. In the `pwrss` package, the default values of the `pwrss.t.2means()` function arguments are set to facilitate standardized definition.

Example: A researcher is trying to find out whether female students studying at universities during COVID-19 have higher levels of depression compared to boys. It is known from previous studies that the standard deviation of depression scores in the total sample (girls + boys) is 6.75 (Hisli, 1989). Again, assuming that the minimum meaningful difference is two units (26 for girls and 24 for boys), the statistical power with 50 participants in each group, or the sample size for 80% power is calculated via defining `alternative="greater"` as

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, kappa = 1,
               alpha = 0.05, n = 50, alternative = "greater")
# Difference between two means
# (independent samples t test)
# H0: mu1 = mu2
# HA: mu1 > mu2
# ------------------------------
#  Statistical power = 0.431
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "greater"
```

```
# Degrees of freedom = 98
# Non-centrality parameter = 1.481
# Type I error rate = 0.05
# Type II error rate = 0.569

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, kappa = 1,
               alpha = 0.05, power = 0.8, alternative = "greater")
# Difference between two means
# (independent samples t test)
# H0: mu1 = mu2
# HA: mu1 > mu2
# ------------------------------
#  Statistical power = 0.8
#  n1 = 142
#  n2 = 142
# ------------------------------
# Alternative = "greater"
# Degrees of freedom = 281.2
# Non-centrality parameter = 2.493
# Type I error rate = 0.05
# Type II error rate = 0.2
```

`The kappa=1` argument refers to the ratio of female participants to male participants (`n1/n2`). `The sd1` argument represents the standard deviation of the first group as well as the pooled standard deviation of the combined data because when the standard deviation of two groups is equal, the pooled standard deviation is similar to the standard deviation of one of the groups (by default, the standard deviation of the second group is equal to the standard deviation of the first group). Statistical power was found to be 43.1%. The hypothesis test can be conducted with a power rate of 80% when there are 142 participants in each group.

Two-way hypothesis testing: Two-way hypothesis testing is used if one believes that the mean of one group in the population (μ_1) is not equal to the mean of the other group (μ_2). The null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_1$

The test statistic is the same as Equation 50. In contrast, since a two-tailed hypothesis test is carried out, one should specify $\alpha/2$ for the type I error rate. Statistical power is calculated as

$$t_k = \Phi_t^{-1}(\alpha/2, v; 0) \tag{56}$$

$$1 - \beta = 1 - \Phi_t(t_k, v; t) + \Phi_t(-t_k, v; t) \quad (57)$$

The minimum required sample size can be found from

$$t_{\alpha/2} + t_\beta = \Phi_z^{-1}(\alpha/2, v; 0) + \Phi_z^{-1}(\beta, v; 0) \quad (58)$$

$$n_2 = \left(t_{\alpha/2} + t_\beta\right)^2 \left(\frac{\frac{\hat{\sigma}_1^{\,2}}{\kappa} + \hat{\sigma}_2^{\,2}}{(\hat{\mu}_1 - \hat{\mu}_2)^2} \right) \quad (59)$$

Example: A researcher is trying to find out whether girls' depression levels are equal to boys'. Suppose the smallest meaningful difference is two units (26 for girls and 24 for boys). Statistical power with 50 participants in each group or the sample size for a power rate of 80% is calculated by defining `alternative="not equal"` as

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, kappa = 1,
               alpha = 0.05, n = 50, alternative = "not equal")
# Difference between two means
#(independent samples t test)
# H0: mu1 = mu2
# HA: mu1 != mu2
# ------------------------------
#  Statistical power = 0.311
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = “not equal”
# Degrees of freedom = 98
# Non-centrality parameter = 1.481
# Type I error rate = 0.05
# Type II error rate = 0.689

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, kappa = 1,
               alpha = 0.05, power = 0.8, alternative = "not equal")
# Difference between two means
# (independent samples t test)
# H0: mu1 = mu2
# HA: mu1 != mu2
# ------------------------------
#  Statistical power = 0.8
#  n1 = 180
#  n2 = 180
# ------------------------------
# Alternative = “not equal”
```

```
# Degrees of freedom = 357.56
# Non-centrality parameter = 2.809
# Type I error rate = 0.05
# Type II error rate = 0.2
```

The power rate with 50 participants in each group is 31.1%. Furthermore, at least 179 participants in each group are needed to conduct the hypothesis test with 80% power.

Non-inferiority or superiority hypothesis testing: In the non-inferiority hypothesis test, the difference between the means of the two groups in the population ($\mu_1 - \mu_2$) is considered to be greater than the margin (δ) on the left side of 0 (negative) or smaller than the margin (δ) on the right side of 0 (positive). In cases where higher values of a variable indicate a positive phenomenon (e.g., achievement test score), the margin is usually negative. The null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $\mu_1 - \mu_2 \leq \delta$

H_A: $\mu_1 - \mu_2 > \delta$

In cases where higher values of the variable indicate a negative phenomenon (e.g., depression score), the margin is usually positive. The null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $\mu_1 - \mu_2 \geq \delta$

H_A: $\mu_1 - \mu_2 < \delta$

In the superiority hypothesis test, the difference in the mean of the two groups in the population ($\mu_1 - \mu_2$) is thought to be smaller than the margin (δ) on the left side of 0 (negative) or greater than the margin (δ) on the right side of 0 (positive). In cases where higher values of the variable indicate a positive phenomenon (e.g., achievement test score), the margin is usually positive. The null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $\mu_1 - \mu_2 \leq \delta$

H_A: $\mu_1 - \mu_2 > \delta$

In cases where higher values of the variable indicate a negative phenomenon (e.g., depression score), the margin is usually negative, and the null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $\mu_1 - \mu_2 \geq \delta$

H_A: $\mu_1 - \mu_2 < \delta$

For both non-inferiority and superiority tests, the test statistics is calculated as

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2 - \delta}{SE(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{\hat{\mu}_1 - \hat{\mu}_2 - \delta}{\sqrt{\frac{\hat{\sigma}_1^{\,2}}{n_1} + \frac{\hat{\sigma}_2^{\,2}}{n_2}}} \tag{60}$$

Equations 52 and 53 are used to calculate the statistical power. The minimum required sample size is calculated as

$$n_2 = (t_\alpha + t_\beta)^2 \left(\frac{\frac{\hat{\sigma}_1^{\,2}}{\kappa} + \hat{\sigma}_2^{\,2}}{(\hat{\mu}_1 - \hat{\mu}_2 - \delta)^2} \right) \tag{61}$$

Example: A researcher is trying to determine whether female students' psychological resilience levels in secondary schools are higher than male students' during COVID-19. Even if the difference between the two means is -1, they will reject the null and conclude that the mean of the first group is higher. Statistical power with 50 participants from each group or the minimum required sample size for a power rate of 80% is calculated by defining `margin=-1` ve `alternative="non-inferior"` as

```
pwrss.t.2means(mu1 = 50, mu2 = 48, sd1 = 7.59, margin = -1,
               alpha = 0.05, n = 50, alternative = "non-inferior")
# Difference between two means
# (independent samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.625
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "non-inferior"
# Degrees of freedom = 98
# Non-centrality parameter = 1.976
# Type I error rate = 0.05
# Type II error rate = 0.375

pwrss.t.2means(mu1 = 50, mu2 = 48, sd1 = 7.59, margin = -1,
               alpha = 0.05, power = 0.8, alternative = "non-inferior")
# Difference between two means
# (independent samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.8
#  n1 = 80
#  n2 = 80
```

```
# ------------------------------
# Alternative = "non-inferior"
# Degrees of freedom = 157.82
# Non-centrality parameter = 2.498
# Type I error rate = 0.05
# Type II error rate = 0.2
```

With 50 participants in each group, the non-inferiority hypothesis testing can be performed with a power rate of 62.5%. Nonetheless, at least 80 participants in each group are required to conduct the hypothesis test with a power rate of 80%.

In superiority hypothesis testing, the researcher would believe that the level of psychological resilience of female students is practically higher than that of male students. The difference between the two means should be greater than 1 to reject null and conclude that the mean of the first group is higher. With 50 participants from each group, the statistical power is 16.1%. At least 714 participants in each group are needed to conduct the superiority test with 80% power rate.

```
pwrss.t.2means(mu1 = 50, mu2 = 48, sd1 = 7.59, margin = 1,
               alpha = 0.05, n = 50, alternative = "superior")
# Difference between two means
# (independent samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.161
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "superior"
# Degrees of freedom = 98
# Non-centrality parameter = 0.659
# Type I error rate = 0.05
# Type II error rate = 0.839

pwrss.t.2means(mu1 = 50, mu2 = 48, sd1 = 7.59, margin = 1,
               alpha = 0.05, power = 0.80, alternative = "superior")
# Difference between two means
# (independent samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.8
#  n1 = 714
#  n2 = 714
```

```
# ----------------------------
# Alternative = "superior"
# Degrees of freedom = 1424.16
# Non-centrality parameter = 2.488
# Type I error rate = 0.05
# Type II error rate = 0.2
```

Equivalence hypothesis testing: In equivalence hypothesis testing, one believes that the absolute value of the difference between the means of the two groups in the population ($|\mu_1 - \mu_2|$) is smaller than some margin (δ). The null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $|\mu_1 - \mu_2| \geq \delta$

H_A: $|\mu_1 - \mu_2| < \delta$

The test statistics can be calculated as

$$t = \frac{|\hat{\mu}_1 - \hat{\mu}_2| - \delta}{SE(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{|\hat{\mu}_1 - \hat{\mu}_2| - \delta}{\sqrt{\frac{\hat{\sigma}_1^{\,2}}{n_1} + \frac{\hat{\sigma}_2^{\,2}}{n_2}}} \tag{62}$$

Statistical power is calculated as

$$t_k = \Phi_t^{-1}(\alpha, v;\ 0) \tag{63}$$

$$1 - \beta = 2\big(1 - \Phi_t(t_k, v; t)\big) - 1 \tag{64}$$

and the minimum required sample size is calculated as

$$t_\alpha + t_{\beta/2} = \Phi_t^{-1}(\alpha, v; 0) + \Phi_t^{-1}(\beta/2, v; 0) \tag{65}$$

$$n_2 = \big(t_\alpha + t_{\beta/2}\big)^2 \left(\frac{\frac{\hat{\sigma}_1^{\,2}}{\kappa} + \hat{\sigma}_2^{\,2}}{(|\hat{\mu}_1 - \hat{\mu}_2| - \delta)^2} \right) \tag{66}$$

Example: A researcher tries to find that female students' psychological resilience levels are equivalent to male students'. They will reject null and claim equivalence when the difference between the two means is less than 1 or greater than -1. The power rate cannot be calculated with 50 participants in each group. The sample size for a power rate of 80% is calculated by defining `margin=1` ve `alternative="equivalent"` as

```
pwrss.t.2means(mu1 = 50, mu2 = 48, sd1 = 7.59, margin = 1,
               alpha = 0.05, n2 = 50, alternative = "equivalent")
# Error: design is not feasible

pwrss.t.2means(mu1 = 50, mu2 = 48, sd1 = 7.59, margin = 1,
```

```
                alpha = 0.05, power = 0.80, alternative = "equivalent")
# Difference between two means
# (independent samples t test)
# H0: |mu1 - mu2| >= margin
# HA: |mu1 - mu2| < margin
# ------------------------------
#  Statistical power = 0.8
#  n1 = 988
#  n2 = 988
# ------------------------------
# Alternative = "equivalent"
# Degrees of freedom = 1973
# Non-centrality parameter = 2.928
# Type I error rate = 0.05
# Type II error rate = 0.2
```

At least 988 participants are needed in each group to test equivalence with a power rate of 80%.

Dependent samples (matched pairs) t-test: Assume that a random variable measured at the first time point (X_1) has a mean of μ_1 and a variance of σ_1^2. Further, assume that the same variable is measured at a second time point (X_2) and it has a mean of μ_2 and a variance of σ_2^2. X_1 and X_2 are dependent measures because they are nested within the person.

Let $x_{11}, x_{12}, x_{13}, \ldots, x_{1n}$ be observed values for X_1 at the first time point and $x_{21}, x_{22}, x_{23}, \ldots, x_{2n}$ are observed values at the second time point. Their means are estimated as

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i} \tag{67}$$

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{2i} \tag{68}$$

and their variances are estimated as

$$\hat{\sigma}_1{}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_{1i} - \hat{\mu}_1)^2 \tag{69}$$

$$\hat{\sigma}_2{}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_{2i} - \hat{\mu}_2)^2 \tag{70}$$

The standard error of the difference $(\hat{\mu}_1 - \hat{\mu}_2)$ is calculated with

$$SE(\hat{\mu}_1 - \hat{\mu}_2) = \sqrt{\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{n}} \quad (71)$$

where r_{12} is the correlation between X_1 and X_2 and defined as

$$r_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1\hat{\sigma}_2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_{1i} - \hat{\mu}_1)(x_{2i} - \hat{\mu}_2)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{1i} - \hat{\mu}_1)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{2i} - \hat{\mu}_2)^2}} \quad (72)$$

One-way hypothesis testing: One-way hypothesis testing is conducted if one believes that the mean of a variable at the first time point is less or greater than the mean of the same variable at the second time point. The null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $\mu_1 \geq \mu_2$ (or $\mu_1 \leq \mu_2$)

H_A: $\mu_1 < \mu_2$ (or $\mu_1 > \mu_2$)

Test statistics is calculated as

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{SE(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{n}}} \quad (73)$$

The statistical power is calculated as in Equations 52 and 53 by defining the degree of freedom as $v = n - 1$. The minimum required sample size is calculated as

$$n = \left(t_\alpha + t_\beta\right)^2 \left(\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{(\hat{\mu}_1 - \hat{\mu}_2)^2}\right) \quad (74)$$

When Cohen's d is used as the standardized mean difference, the statistical power or the minimum required sample size is calculated by defining $\hat{\mu}_1 = d$, $\hat{\mu}_2 = 0$, $\hat{\sigma}_1{}^2 = \sqrt{1/2(1 - r_{12})}$ ve $\hat{\sigma}_2{}^2 = \sqrt{1/2(1 - r_{12})}$.

Example: A researcher is planning to organize a one-week program based on cognitive behavioral psychotherapy to reduce the depression levels of students studying at universities during the COVID-19 pandemic. They will measure participants' depression levels at the beginning and end of the program (pretest and posttest). Bulus and Koyuncu (2021) found that the pretest for non-cognitive outcomes in the field of psychological counseling explained 0.29 of the variance in the posttest scores ($r_{12}^2 = 0.29$). This coefficient of determination shows that the correlation between the pretest and the posttest scores is $r_{12} = \sqrt{0{,}29} = 0{,}54$. It is also known from previous studies that the standard deviation of depression scores was around 6.75 (Hisli, 1989). Assuming that the minimum noteworthy reduction in depression symptoms is two units (pretest mean of 26 and posttest mean of 24 points), statistical power with 50 participants or the minimum required sample size for a power rate of 80% is found by defining `paired=TRUE` ve `alternative="greater"` as

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, n = 50,
               paired = TRUE, paired.r = 0.54, alternative = "greater")
# Difference between two means
# (paired samples t test)
# H0: mu1 = mu2
# ha: mu1 > mu2
# ------------------------------
#  Statistical power = 0.695
#  n = 50
# ------------------------------
# Alternative = "greater"
# Degrees of freedom = 49
# Non-centrality parameter = 2.184
# Type I error rate = 0.05
# Type II error rate = 0.305

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, power = 0.8,
               paired = TRUE, paired.r = 0.54, alternative = "greater")
# Difference between two means
# (paired samples t test)
# H0: mu1 = mu2
# ha: mu1 > mu2
# ------------------------------
#  Statistical power = 0.8
#  n = 67
# ------------------------------
# Alternative = "greater"
# Degrees of freedom = 65.31
# Non-centrality parameter = 2.516
# Type I error rate = 0.05
# Type II error rate = 0.2
```

As mentioned earlier, the `sd1` argument can represent the standard deviation of the first group as well as the pooled standard deviation of the combined data because, by default, the standard deviation of the second group is equal to the standard deviation of the first group. According to this calculation, the hypothesis test can be conducted with 69.5% power. At least 67 participants are needed to achieve a power rate of 80%. If an increase in the level of depression is expected, that is, to find that the mean of pretest scores was lower than the mean of the posttest scores, then the argument `alternative="less"` should be used.

Two-way hypothesis testing: Two-way hypothesis testing is used when one believes that the mean of a variable in the population in the first time point (μ_1) is different from the mean in the second time point (μ_2). The null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_1$

The test statistic is similar to that in Equation 73. In contrast, since $\hat{\mu}_1$ could be higher as well as lower than $\hat{\mu}_2$, the type I error rate is set to $\alpha/2$. The statistical power is calculated as in Equations 56 and 57 by defining the degree of freedom as $v = n - 1$. After modifying the type I error rate and calculating degrees of freedom the sample size can be calculated from Equation 54 as

$$n = (t_{\alpha/2} + t_{\beta})^2 \left(\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{(\hat{\mu}_1 - \hat{\mu}_2)^2} \right) \tag{75}$$

Example: Reconsider the previous example, but now the level of depression may decrease or increase. The statistical power with 50 participants or the minimum required sample size for a power rate of 80% is found by defining `alternative="not equal"` as

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, n = 50,
               paired = TRUE, paired.r = 0.54, alternative = "not equal")
# Difference between two means
# (paired samples t test)
# H0: mu1 = mu2
# HA: mu1 != mu2
# ------------------------------
#  Statistical power = 0.572
#  n = 50
# ------------------------------
# Alternative = "not equal"
# Degrees of freedom = 49
# Non-centrality parameter = 2.184
# Type I error rate = 0.05
# Type II error rate = 0.428

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, power = 0.8,
               paired = TRUE, paired.r = 0.54, alternative = "not equal")
# Difference between two means
# (paired samples t test)
# H0: mu1 = mu2
# HA: mu1 != mu2
# ------------------------------
#  Statistical power = 0.8
#  n = 85
# ------------------------------
# Alternative = "not equal"
# Degrees of freedom = 83.21
# Non-centrality parameter = 2.835
```

```
# Type I error rate = 0.05
# Type II error rate = 0.2
```

The power rate is 57.2% with 50 participants. At least 85 participants are needed to achieve a power rate of 80%.

Non-inferiority or superiority hypothesis testing: Non-inferiority hypothesis testing is used when one believes that the mean difference between the first and second time point ($\mu_1 - \mu_2$) is greater than the margin (δ) when it is on the left side of 0 (negative) or less than the margin when it is on the right side of 0 (positive). In cases where higher values of a variable indicate a positive phenomenon (e.g., achievement test score), the margin is usually negative, and the null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $\mu_1 - \mu_2 \leq \delta$

H_A: $\mu_1 - \mu_2 > \delta$

When higher values of the variable indicate a negative phenomenon (e.g., depression score), the margin is usually positive, and the null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $\mu_1 - \mu_2 \geq \delta$

H_A: $\mu_1 - \mu_2 < \delta$

Superiority hypothesis testing is used when one believes that the mean difference between the first and second time points ($\mu_1 - \mu_2$) is less than the margin (δ) when it is on the left side of 0 (negative) or greater than the margin when it is on the right side of 0 (positive). In cases where higher values of a variable indicate a positive phenomenon (e.g., achievement test score), the margin is usually positive, and the null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $\mu_1 - \mu_2 \leq \delta$

H_A: $\mu_1 - \mu_2 > \delta$

When higher values of the variable indicate a negative phenomenon, the margin is usually negative, and the null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $\mu_1 - \mu_2 \geq \delta$

H_A: $\mu_1 - \mu_2 < \delta$

The test statistic is calculated as

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2 - \delta}{SE(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{\hat{\mu}_1 - \hat{\mu}_2 - \delta}{\sqrt{\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{n}}} \quad (76)$$

The statistical power is calculated using Equations 52 and 53 by defining the degree of freedom as $v = n - 1$. After specifying the degrees of freedom in Equation 54, the sample size is calculated as

$$n = (t_\alpha + t_\beta)^2 \left(\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{(\hat{\mu}_1 - \hat{\mu}_2 - \delta)^2} \right) \tag{77}$$

Example: Revisiting the previous example, but now assuming that the researcher would conclude that there is a decrease in the level of depression as long as the difference is greater than -1. Statistical power with 50 participants or the minimum required sample size for a power rate of 80% is calculated by defining `margin=-1` ve `alternative="non-inferior"` as

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, n = 50,
               paired = TRUE, paired.r = 0.54,
               alternative = "non-inferior", margin = -1)
# Difference between two means
# (paired samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.944
#  n = 50
# ------------------------------
# Alternative = "non-inferior"
# Degrees of freedom = 49
# Non-centrality parameter = 3.276
# Type I error rate = 0.05
# Type II error rate = 0.056

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, power = 0.8,
               paired = TRUE, paired.r = 0.54,
               alternative = "non-inferior", margin = -1)
# Difference between two means
# (paired samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.8
#  n = 31
# ------------------------------
# Alternative = "non-inferior"
# Degrees of freedom = 29.34
# Non-centrality parameter = 2.552
# Type I error rate = 0.05
# Type II error rate = 0.2
```

In conclusion, with 50 participants, considering a margin of -1, the difference of two units between the mean pretest and mean posttest scores can be detected with 94.4% power rate. Only 31 participants are needed for a power rate 80%.

Suppose that the researcher would conclude that there is a decrease in the level of depression as long as the difference is greater than 1. Statistical power with 50 participants or the minimum required sample size for a power rate of 80% is calculated by defining `margin=1` ve `alternative="superior"` as

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, n = 50,
               paired = TRUE, paired.r = 0.54,
               alternative = "superior", margin = 1)
# Difference between two means
# (paired samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.285
#  n = 50
# ------------------------------
# Alternative = "superior"
# Degrees of freedom = 49
# Non-centrality parameter = 1.092
# Type I error rate = 0.05
# Type II error rate = 0.715

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, power = 0.8,
               paired = TRUE, paired.r = 0.54,
               alternative = "superior", margin = 1)
# Difference between two means
# (paired samples t test)
# H0: mu1 - mu2 <= margin
# HA: mu1 - mu2 > margin
# ------------------------------
#  Statistical power = 0.8
#  n = 261
# ------------------------------
# Alternative = "superior"
# Degrees of freedom = 259.67
# Non-centrality parameter = 2.494
# Type I error rate = 0.05
# Type II error rate = 0.2
```

With 50 participants, the superiority test has a power rate of 28.5%. At least 261 participants are needed to perform the test with 80% power rate.

Equivalence hypothesis testing: In equivalence hypothesis testing, one believes that there is no difference between the means of two time points while considering a margin. The null hypothesis is rejected when the difference is greater than the margin on the left side of 0 (negative) and smaller than the margin on the right side of 0 (positive). The null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $|\mu_1 - \mu_2| \geq \delta$

H_A: $|\mu_1 - \mu_2| < \delta$

The test statistics is calculated as

$$t = \frac{|\hat{\mu}_1 - \hat{\mu}_2| - \delta}{SE(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{|\hat{\mu}_1 - \hat{\mu}_2| - \delta}{\sqrt{\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{n}}} \tag{78}$$

Equations 63 and 64 are used to calculate the statistical power. After plugging $v = n - 1$ in Equation 65, the sample size is calculated as

$$n = \left(t_\alpha + t_{\beta/2}\right)^2 \left(\frac{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2 - 2\hat{\sigma}_1\hat{\sigma}_2 r_{12}}{(|\hat{\mu}_1 - \hat{\mu}_2| - \delta)^2}\right) \tag{79}$$

Example: Revisiting the previous example, assume that the researcher will conclude that there is no change in the level of depression as long as the absolute value of the difference is smaller than 1. Statistical power with 50 participants or the minimum required sample size for a power rate of 80% is calculated by defining `margin=1` ve `alternative="equivalent"` as

```
pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, n = 50,
               paired = TRUE, paired.r = 0.54,
               alternative = "equivalent", margin = 1)
# Error: design is not feasible

pwrss.t.2means(mu1 = 26, mu2 = 24, sd1 = 6.75, alpha = 0.05, power = 0.8,
               paired = TRUE, paired.r = 0.54,
               alternative = "equivalent", margin = 1)
# Difference between two means
# (paired samples t test)
# H0: |mu1 - mu2| >= margin
# HA: |mu1 - mu2| < margin
# ------------------------------
#  Statistical power = 0.8
#  n = 361
```

```
# ------------------------------
# Alternative = "equivalent"
# Degrees of freedom = 359.59
# Non-centrality parameter = 2.933
# Type I error rate = 0.05
# Type II error rate = 0.2
```

It is not possible to calculate the power rate with 50 participants. At least 361 participants are needed to perform the equivalence test with 80% power rate.

Comparing Pearson's Correlation against a Constant

The Fisher transformation is used to compare a correlation ($\hat{r}$) against a constant (r_0) (Cohen, 1988). The transformed values can be obtained via

$$\hat{z} = \frac{1}{2}\log\left(\frac{1+\hat{r}}{1-\hat{r}}\right) \tag{80}$$

$$z_0 = \frac{1}{2}\log\left(\frac{1+r_0}{1-r_0}\right) \tag{81}$$

The test statistic for the z-test is

$$z = \frac{\hat{z} - z_0}{\sqrt{\frac{1}{n-3}}} \tag{82}$$

where n is the sample size.

One-way hypothesis testing: One-way hypothesis testing is used when one believes that a correlation between two variables in the population (r) is smaller or greater than a fixed value (r_0). The null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $r \geq r_0$ (or $r \leq r_0$)

H_A: $r < r_0$ (or $r > r_0$)

Statistical power is calculated using Equations 7 and 8. The sample size is calculated as

$$n = \frac{\left(z_\alpha + z_\beta\right)^2}{(\hat{z} - z_0)^2} + 3 \tag{83}$$

In `pwrss` R package, $r_0 = 0$ is by default but can be modified by the user. Thus, a researcher trying to find out whether a correlation is greater than, less than, or not equal to zero does not need to specify r_0.

Example: A researcher is trying to find whether there is a positive relationship between a collaborative learning environment and school belonging. A previous study found a correlation of 0.24 between the

collaborative learning environment and school belonging (Ozcan and Bulus, 2022). If there is no evidence of an expected correlation value, the classifications of Cohen (1988) or Gignac and Szodorai (2016) can be used. Statistical power with 50 participants or the minimum required sample size for a power rate of 80% is calculated by defining `alternative="greater"` as

```
pwrss.z.corr(r = .24, n = 50, alternative = "greater")
# One correlation compared to a constant
# (one sample z test)
# H0: r = r0
# HA: r > r0
# ------------------------------
#  Statistical power = 0.513
#  n = 50
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 1.678
# Type I error rate = 0.05
# Type II error rate = 0.487

pwrss.z.corr(r = .24, power = .8, alternative = "greater")
# One correlation compared to a constant
3 (one sample z test)
# H0: r = r0
# HA: r > r0
# ------------------------------
#  Statistical power = 0.8
#  n = 107
# ------------------------------
# Alternative = "greater"
# Non-centrality parameter = 2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

Statistical power with 50 participants is 51.3%. At least 107 participants are needed to perform the hypothesis test with a power rate of 80%.

Two-way hypothesis testing: Two-way hypothesis testing is used when one believes that a correlation in the population (r) is different from a fixed correlation value (r_0). The null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $r = r_0$

H_A: $r \neq r_0$

Statistical power is calculated as in Equations 11 and 12. The sample size is calculated as

$$n = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2}{(\hat{z} - z_0)^2} + 3 \tag{84}$$

Example: Revisiting the previous example, suppose the researcher is now trying to find out whether the relationship between the collaborative learning environment and school belonging is different from 0. Statistical power with 50 participants or the minimum required sample size for a power rate of 80% is calculated by defining `alternative="not equal"` as

```
pwrss.z.corr(r = .24, n = 50, alternative = "not equal")
# One correlation compared to a constant
# (one sample z test)
# H0: r = r0
# HA: r != r0
# ------------------------------
#  Statistical power = 0.389
#  n = 50
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 1.678
# Type I error rate = 0.05
# Type II error rate = 0.611

pwrss.z.corr(r = .24, power = .8, alternative = "not equal")
# One correlation compared to a constant
# (one sample z test)
# H0: r = r0
# HA: r != r0
# ------------------------------
#  Statistical power = 0.8
#  n = 135
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = 2.802
# Type I error rate = 0.05
# Type II error rate = 0.2
```

Statistical power with 50 participants is 38.9%. At least 135 participants are needed to perform the hypothesis test with a power rate of 80%.

Comparing Two Pearson Correlations

Let $\hat{r}_1$ and $\hat{r}_2$ be two correlations in samples with size n_1 and n_2, respectively. Fisher transformation is used to compare and test the two correlation coefficients (Cohen, 1988):

$$\hat{z}_1 = \frac{1}{2}\log\left(\frac{1+\hat{r}_1}{1-\hat{r}_1}\right) \tag{85}$$

$$\hat{z}_2 = \frac{1}{2}\log\left(\frac{1+\hat{r}_2}{1-\hat{r}_2}\right) \tag{86}$$

The test statistics is

$$z = \frac{\hat{z}_1 - \hat{z}_2}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}} \tag{87}$$

One-way hypothesis testing: One-way hypothesis testing is used when one believes that one of the correlations is less or greater than the other in the population. The null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $r_1 \geq r_2$ (or $r_1 \leq r_2$)

H_A: $r_1 < r_2$ (or $r_1 > r_2$)

Statistical power is calculated using Equations 7 and 8. The minimum required sample size for the second group is found by feeding

$$f(n_2) = \left(\frac{1}{\kappa n_2 - 3} + \frac{1}{n_2 - 3}\right) - \frac{(z_1 - z_2)^2}{\left(z_\alpha + z_\beta\right)^2} = 0 \tag{88}$$

expression into the `uniroot()` function in R. Sample size for the first group is found from $n_1 = \kappa n_2$.

Example: It is known from previous studies that there is a positive relationship between collaborative learning environment and school belonging in individualistic societies ($r_1 = 0.23$) and in collectivist societies ($r_2 = 0.25$) (Ozcan and Bulus, 2022). Assuming that a difference of 0.02 in correlation coefficients is significant in practice and that the correlation in individualistic communities is smaller, statistical power with 50 participants or the minimum required sample size for a power rate of 80% is calculated by defining `alternative="less"` as

```
pwrss.z.2corrs(r1 = .23, r2 = 0.25,
               n2 = 50, alternative = "less")
# Difference between two correlations
3 (independent samples z test)
# H0: r1 = r2
# HA: r1 < r2
# ------------------------------
#  Statistical power = 0.062
#  n1 = 50
#  n2 = 50
```

```
# ------------------------------
# Alternative = "less"
# Non-centrality parameter = -0.103
# Type I error rate = 0.05
# Type II error rate = 0.938

pwrss.z.2corrs(r1 = .23, r2 = 0.25,
               power = 0.8, alternative = "less")
# Difference between two correlations
# (independent samples z test)
# H0: r1 = r2
# HA: r1 < r2
# ------------------------------
#  Statistical power = 0.8
#  n1 = 27455
#  n2 = 27455
# ------------------------------
# Alternative = "less"
# Non-centrality parameter = -2.486
# Type I error rate = 0.05
# Type II error rate = 0.2
```

Statistical power with 50 participants in each group is 6.2%. At least 27455 participants are needed in each group to perform the hypothesis test with a power rate of 80%.

Two-way hypothesis testing: Two-way hypothesis testing is used when one believes that two correlations are not equal. The null (H_0) and alternative (H_A) hypotheses are constructed as follows:

H_0: $r_1 = r_2$

H_A: $r_1 \neq r_2$

Statistical power is calculated using Equations 11 and 12. The minimum required sample size in the second group is found from

$$f(n_2) = \left(\frac{1}{\kappa n_2 - 3} + \frac{1}{n_2 - 3}\right) - \frac{(z_1 - z_2)^2}{\left(z_{\alpha/2} + z_\beta\right)^2} = 0 \qquad (90)$$

by feeding the expression into `uniroot()` function in the R package. The minimum required sample size for the first group is found from $n_1 = \kappa n_2$.

Example: Revisiting the previous example, suppose that the difference of 0.02 units between the two correlations can be negative or positive. Statistical power with 50 participants in each group or the minimum required sample size for a power rate of 80% is calculated by defining `alternative="not equal"` as

```
pwrss.z.2corrs(r1 = .23, r2 = 0.25,
               n2 = 50, alternative = "not equal")
# Difference between two correlations
# (independent samples z test)
# H0: r1 = r2
# HA: r1 != r2
# ------------------------------
#  Statistical power = 0.051
#  n1 = 50
#  n2 = 50
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = -0.103
# Type I error rate = 0.05
# Type II error rate = 0.949

pwrss.z.2corrs(r1 = .23, r2 = 0.25,
               power = 0.8, alternative = "not equal")
# Difference between two correlations
# (independent samples z test)
# H0: r1 = r2
# HA: r1 != r2
# ------------------------------
#  Statistical power = 0.8
#  n1 = 34854
#  n2 = 34854
# ------------------------------
# Alternative = "not equal"
# Non-centrality parameter = -2.802
# Type I error rate = 0.05
# Type II error rate = 0.2
```

Statistical power with 50 participants in each group is 5.1%. At least 34854 participants in each group are needed to perform the hypothesis test with a power rate of 80%.

R^2 (or ΔR^2) in Multiple Linear Regression

One may want to add all variables of interest to the regression model and investigate the proportion of variance explained in the predicted variable (R^2), or one may add variables in sets, one step at a time (hierarchical regression analysis), and investigate the change in the variance explained (ΔR^2). The null (H_0) and alternative (H_A) hypotheses are formed as

H_0: $\Delta R^2 = 0$

H_A: $\Delta R^2 > 0$

The ratio of explained variance to the unexplained variance is stated in terms of Cohen's f^2 (Cohen, 1988) as

$$f^2 = \frac{\Delta R^2}{1 - \Delta R^2} \tag{91}$$

Presume one is interested in m set of predictors out of a total of k predictors in a sample of n observations. The test statistics is stated as

$$\lambda = f^2 n \tag{92}$$

When the null hypothesis is correct, the test statistic follows the central F distribution with $u = m$ degrees of freedom for the numerator and $v = n - k - 1$ degrees of freedom for the denominator. When the alternative hypothesis is correct, the test statistic follows the F distribution with centrality parameter λ, with the same degrees of freedom. For the F distribution, denoting the cumulative distribution function with Φ_F, and the inverse cumulative distribution function with Φ_F^{-1}, the statistical power is calculated as

$$F_k = \Phi_F^{-1}(\alpha, u, v;\ 0) \tag{93}$$

$$1 - \beta = \Phi_F(F_k, u, v;\ \lambda) \tag{94}$$

To calculate the sample size, Equation 94 is revised and written as follows:

$$f(n) = \Phi_F(F_k, u, v;\ \lambda) + \beta - 1 = 0 \tag{95}$$

The sample size corresponding to the power rate is determined using the `uniroot()` function in R program.

Example: A researcher wants to find out the effect of certain defense mechanisms during the COVID-19 pandemic (talking to friends, exercising, social media, reading books, hobbies, religious activities, alcohol, and researching about COVID) on psychosomatic symptoms, beyond demographics (age and gender). Previous studies on the same topic found that only when age and sex were added to the regression model (m = 2), only 0.01% of the variance in psychosomatic symptoms was explained. In contrast, when eight defense mechanisms were added along with age and sex (k = 8 + 2), 24% of the variance in psychosomatic symptoms was explained (Otanga et al., 2022). When eight variables are added to the model, the change is $\Delta R^2 = 0.024 - 0.01 = 0.23$. Statistical power with 50 participants or the sample size for a power rate of 80% is

```
pwrss.f.reg(k  = 10, m = 8, n  = 50, r2 = 0.23)
# R-squared change in hierarchical linear regression (F test)
# H0: r2 = 0
# HA: r2 > 0
```

```
# ------------------------------
#  Statistical power = 0.701
#  n = 50
# ------------------------------
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 39
# Non-centrality parameter = 14.935
# Type I error rate = 0.05
# Type II error rate = 0.299

pwrss.f.reg(k = 10, m = 8, power = 0.8, r2 = 0.23)
# R-squared change in hierarchical linear regression (F test)
# H0: r2 = 0
# HA: r2 > 0
# ------------------------------
#  Statistical power = 0.8
#  n = 59
# ------------------------------
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 25.385
# Non-centrality parameter = 10.868
# Type I error rate = 0.05
# Type II error rate = 0.2
```

Statistical power with 50 participants is 70.1%. At least 59 participants are needed to perform the hypothesis test with a power rate of 80%. If a researcher is interested in all of the predictors, they will need to run the following code to find the sample size

```
pwrss.f.reg(k = 10, power = 0.8, r2 = 0.24)
# R-squared compared to 0 in linear regression (F test)
# H0: r2 = 0
# HA: r2 > 0
# ------------------------------
#  Statistical power = 0.8
#  n = 62
# ------------------------------
# Numerator degrees of freedom = 10
# Denominator degrees of freedom = 50.168
# Non-centrality parameter = 19.316
# Type I error rate = 0.05
# Type II error rate = 0.2
```

Analysis of Variance and Covariance (ANOVA and ANCOVA)

One could be interested in testing the difference between at least two group means, for which they can use Analysis of Variance (ANOVA). They may also adjust mean differences for covariates such as pretest, for which they can use Analysis of Co-variance (ANCOVA). One can use Cohen's f^2 or (partial) η^2 as the effect size. The null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $\eta^2 = 0$ (or $f^2 = 0$)

H_A: $\eta^2 > 0$ (or $f^2 > 0$)

The ratio of the explained variance in the outcome by the groups (or group interactions) to the remaining variance is (Cohen, 1988)

$$f^2 = \frac{\eta^2}{1 - \eta^2} \tag{96}$$

Suppose that data are collected from n participants, that the factor consists of k groups, and that the mean differences are corrected for p covariate. The test statistic is calculated as

$$\lambda = f^2 n \tag{97}$$

When the null hypothesis is correct, the test statistic follows the central F distribution with $u = k - 1$ degrees of freedom for the numerator and $v = n - k - p$ degrees of freedom for the denominator. When the alternative hypothesis is correct, the test statistic follows the F distribution with centrality parameter λ, with the same degrees of freedom.

The ANOVA or ANCOVA model may consist of more than one factor (single factor, two-factor, or three-factor), each of these factors may include a different number of groups, and the interactions between these factors may be of interest. The statistical power is calculated using Equations 93 and 94. Equation 95 is fed into the `uniroot()` function in the R program to find the sample size corresponding to the desired power level.

Example: Aslan (2019) applied the ANCOVA test to find the effect of argumentation-based teaching and scenario-based learning methods. They found $\eta^2 = 0.14$ for the groups (two treatment and one control group). Statistical power with 50 participants across all groups or the total sample size for a power rate of 80% is calculated as

```
pwrss.f.ancova(eta2 = 0.14, n = 50,
               n.way = 1, n.levels = 3, n.covariates = 1)
# One-way Analysis of Covariance (ANCOVA)
#  H0: 'eta2' or 'f2' = 0
#  HA: 'eta2' or 'f2' > 0
# ----------------------------------
```

```
# Factor A: 3 levels
# ----------------------------------
# Given eta2 = 0.14 or f2 = 0.163
#  Statistical power = 0.695
# ----------------------------------
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 46
# Non-centrality parameter = 8.14
```

```
pwrss.f.ancova(eta2 = 0.14, power = 0.8,
               n.way = 1, n.levels = 3, n.covariates = 1)
# One-way Analysis of Covariance (ANCOVA)
#  H0: 'eta2' or 'f2' = 0
#  HA: 'eta2' or 'f2' > 0
# ----------------------------------
# Factor A: 3 levels
# ----------------------------------
# Given eta2 = 0.14 or f2 = 0.163
#  Total n = 63
# ----------------------------------
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 58.33
# Non-centrality parameter = 10.15
```

`n.covariates=1` specification indicates that only the pretest was added to the model as a covariate. The `n.levels` argument is used to describe how many levels (groups) the factor consists of, which is specified as three because there were two experiments and one control group. If there is more than one factor, e.g., two-factor ANOVA or ANCOVA, the specification would be `n.levels=c(3,2)`. This specification indicates that there are two factors in the ANOVA or ANCOVA model and that the first factor consists of three levels (groups), and the second factor consists of two levels (groups).

Repeated Measures Analysis of Variance (Repeated *Measures* ANOVA)

Repeated measures ANOVA is used when data are collected from the same subjects at multiple time points (time as within factor) and from two or more groups (as between factor). If one is interested in whether there is a difference between at least two group means while conditioning on time, whether there is a difference between at least two time points while conditioning on the group membership, or whether there is an interaction between group membership and time. Cohen's f^2 or (partial) η^2 can be used as the effect size measure. The null (H_0) and alternative (H_A) hypotheses are formed as follows:

H_0: $\eta^2 = 0$ (or $f^2 = 0$)

H_A: $\eta^2 > 0$ (or $f^2 > 0$)

Suppose that data are collected from n participants, that the group factor has k levels, and the time factor has m points. In this case, the test statistics for the group factor is

$$\lambda = f^2 \left(\frac{m}{1 + (m-1)\rho} \right) n\epsilon \tag{98}$$

The test statistic for the time factor or group x time interaction is

$$\lambda = f^2 \left(\frac{m}{1 - \rho} \right) n\epsilon \tag{99}$$

ϵ is the sphericity correction factor which takes values between $1/(m-1)$ and 1. For testing the group factor, when the null hypothesis is correct, the test statistic follows the central F distribution with $u = k - 1$ degrees of freedom for the numerator and $v = n - k$ degrees of freedom for the denominator. When the alternative hypothesis is correct, the test statistic follows the F distribution with centrality parameter λ (Equation 98), with the same degrees of freedom.

For testing the time factor, when the null hypothesis is correct, the test statistic follows the central F distribution with $u = (m-1)\epsilon$ degrees of freedom for the numerator and $v = (n-k)(m-1)\epsilon$ degrees of freedom for the denominator. When the alternative hypothesis is correct, the test statistic follows the F distribution with centrality parameter λ (Equation 99), with the same degrees of freedom.

For testing the group x time interaction, when the null hypothesis is correct, the test statistic follows the central F distribution with $u = (k-1)(m-1)\epsilon$ degrees of freedom for the numerator and $v = (n-k)(m-1)\epsilon$ degrees of freedom for the denominator. When the alternative hypothesis is correct, the test statistic follows the F distribution with centrality parameter λ (Equation 99), with the same degrees of freedom.

The statistical power is calculated using Equations 93 and 94. Equation 95 is fed into the `uniroot()` function in the R program to calculate the sample size that satisfies the desired power level.

Example: Kartal et al. (2016) created two different experimental groups (in-class play or computer-aided) and a control group aiming to increase the phonological awareness of preschool and first grades and carried out measurements at three time points (pretest, posttest, and follow-up test). Kartal et al. (2016) collected data from 53 participants and found that the effect of the time factor was η_T^2 = 0.56, the effect of the group factor was η_G^2 = 0.47, and the effect of the time x group interaction was η_{TxG}^2 = 0.10.

For the group factor, statistical power with 53 participants or sample size for a power rate of 80% is calculated as

```
##--------------- group effect ---------------- ##
pwrss.f.rmanova(eta2 = 0.47, corr.rm = .5,
                n.levels  = 3, n.rm = 3,
                alpha = 0.05, n = 53, type = "between")
```

```
# One-way repeated measures analysis of variance (F test)
# H0: eta2 = 0 (or f2 = 0)
# HA: eta2 > 0 (or f2 > 0)
# ------------------------------
# Number of levels (groups) = 3
# Number of measurement time points = 3
# ------------------------------
#  Statistical power = 1
#  Total n = 53
# ------------------------------
# Type of the effect = "between"
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 50
# Non-centrality parameter = 70.5
# Type I error rate = 0.05
# Type II error rate = 0

pwrss.f.rmanova(eta2 = 0.47, corr.rm = .5,
                n.levels  = 3, n.rm = 3,
                alpha = 0.05, power = 0.8, type = "between")
# One-way repeated measures analysis of variance (F test)
# H0: eta2 = 0 (or f2 = 0)
# HA: eta2 > 0 (or f2 > 0)
# ------------------------------
# Number of levels (groups) = 3
# Number of measurement time points = 3
# ------------------------------
#  Statistical power = 0.8
#  Total n = 11
# ------------------------------
# Type of the effect = "between"
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 7.871
# Non-centrality parameter = 14.46
# Type I error rate = 0.05
# Type II error rate = 0.2
```

Statistical power with 53 participants is 100%. At least 11 participants are needed to perform the hypothesis test with a power rate of 80%.

For the time factor, statistical power with 53 participants or sample size for a power rate of 80% is calculated as

```
##-------------- time effect --------------- ##
pwrss.f.rmanova(eta2 = 0.56, corr.rm = .5,
```

```
                n.levels  = 3, n.rm = 3,
                alpha = 0.05, n = 53, type = "within")
# One-way repeated measures analysis of variance (F test)
# H0: eta2 = 0 (or f2 = 0)
# HA: eta2 > 0 (or f2 > 0)
# ------------------------------
# Number of levels (groups) = 3
# Number of measurement time points = 3
# ------------------------------
#  Statistical power = 1
#  Total n = 53
# ------------------------------
# Type of the effect = "within"
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 100
# Non-centrality parameter = 404.727
# Type I error rate = 0.05
# Type II error rate = 0

pwrss.f.rmanova(eta2 = 0.56, corr.rm = .5,
                n.levels  = 3, n.rm = 3,
                alpha = 0.05, power = 0.8, type = "within")
# One-way repeated measures analysis of variance (F test)
# H0: eta2 = 0 (or f2 = 0)
# HA: eta2 > 0 (or f2 > 0)
# ------------------------------
# Number of levels (groups) = 3
# Number of measurement time points = 3
# ------------------------------
#  Statistical power = 0.8
#  Total n = 5
# ------------------------------
# Type of the effect = "within"
# Numerator degrees of freedom = 2
# Denominator degrees of freedom = 2.826
# Non-centrality parameter = 33.699
# Type I error rate = 0.05
# Type II error rate = 0.2
```

Statistical power with 53 participants is 100%. At least 5 participants are needed to perform the hypothesis test with a power rate of 80%.

Finally, for the group x time interaction, statistical power with 53 participants or sample size for a power rate of 80% is calculated as

```
##----- Group X Time Interaction Effect ------- ##

pwrss.f.rmanova(eta2 = 0.1, corr.rm = .5,
                n.levels  = 3, n.rm = 3,
                alpha = 0.05, n = 53, type = "interaction")
# One-way repeated measures analysis of variance (F test)
# H0: eta2 = 0 (or f2 = 0)
# HA: eta2 > 0 (or f2 > 0)
# ------------------------------
# Number of levels (groups) = 3
# Number of measurement time points = 3
# ------------------------------
#  Statistical power = 0.999
#  Total n = 53
# ------------------------------
# Type of the effect = "interaction"
# Numerator degrees of freedom = 4
# Denominator degrees of freedom = 100
# Non-centrality parameter = 35.333
# Type I error rate = 0.05
# Type II error rate = 0.001

pwrss.f.rmanova(eta2 = 0.10, corr.rm = .5,
                n.levels  = 3, n.rm = 3,
                alpha = 0.05, power = 0.8, type = "interaction")
# One-way repeated measures analysis of variance (F test)
# H0: eta2 = 0 (or f2 = 0)
# HA: eta2 > 0 (or f2 > 0)
# ------------------------------
# Number of levels (groups) = 3
# Number of measurement time points = 3
# ------------------------------
#  Statistical power = 0.8
#  Total n = 21
# ------------------------------
# Type of the effect="interaction"
# Numerator degrees of freedom = 4
# Denominator degrees of freedom = 34.973
# Non-centrality parameter = 13.658
# Type I error rate = 0.05
# Type II error rate = 0.2
```

Statistical power with 53 participants is 99.9%. At least 21 participants are needed to perform the hypothesis test with a power rate of 80%.

New Approaches in Complex Research Designs

Recently, there has been a rapid increase in power analysis programs. In particular, formulas for calculating power were derived in the areas of experimental, quasi-experimental, and weak experimental designs (simple and multilevel) (Bloom, 1995, 2006, 2012; Bulus, 2022; Bulus and Dong, 2021; Cattaneo, Titiunik, and Vazquez-Bare, 2019; Dong et al., 2021; Hedges and Rhoads, 2010; Kelcey et al., 2017a, 2017b; Konstantopoulos, 2008a, 2008b; Schochet, 2008, 2009; and many others) and these formulas are implemented in Excel files (Dong and Maynard, 2013), in R packages (Bulus and Dong, 2021; Bulus et al., 2021; Cattaneo et al., 2019), and web applications (see Table 1).

In addition, in experimental studies, one could be interested in moderators and mediators of the intervention effect as well as the main effect. To our best knowledge, PowerUp! Excel files (https://www.causalevaluation.org/power-analysis.html) and PowerUpR R package (Bulus et al., 2021) is the only one that comprehensively implements all three design features for experimental designs.

Table 1. *Some Web Based Open Access Power Analysis Programs*

Explanation	Link
General purpose hypothesis testing	https://pwrss.shinyapps.io/index/ https://pwrss.shinyapps.io/lang-en/ https://pwrss.shinyapps.io/lang-tr/
General purpose hypothesis testing	http://biostatapps.inonu.edu.tr/WSSPAS/
General purpose hypothesis testing	http://powerandsamplesize.com/
General purpose hypothesis testing	https://webpower.psychstat.org/wiki/models/index/
Multilevel randomized experimental designs	https://powerupr.shinyapps.io/index/
Multilevel regression discontinuity designs	https://cosa.shinyapps.io/index/
Structural Equation Modeling	https://yilinandrewang.shinyapps.io/pwrSEM/
Mediation Analysis	https://davidakenny.shinyapps.io/MedPower/

Wang and Rhemtulla (2021) developed an R-based pwrSEM application based on Monte-Carlo (MC) simulation to calculate the statistical power for structural parameters in structural equation modeling (SEM). Similarly, the statistical power for any desired parameter in an SEM model can be calculated with the MONTECARLO command in *Mplus* (Muthén and Muthén, 1998-2015) program. For simpler mediation models, one can use David A. Kenny's web application (see Table 1). In simple regression models where the effect of moderator variables is of interest, the InteractionPoweR R program can be useful (Baranger et al., 2022).

Kaynakça

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences* (2nd ed.). Routledge.

Arslan, G. (2015). Ergenlerde psikolojik sağlamlık: Bireysel koruyucu faktörlerin rolü. *Turkish Psychological Counseling and Guidance Journal*, *5*(44), 73-82. https://dergipark.org.tr/tr/download/article-file/631450

Arslan, A. K., Yasar, S, Colak, C., & Yologlu, S. (2018). WSSPAS: An interactive web application for sample size and power analysis with R using Shiny. *Türkiye Klinikleri Biyoistatistik Dergisi*, *10*(3), 224-246. http://doi.org/10.5336/biostatic.2018-62787

Aslan, S. (2019). The ımpact of argumentation-based teaching and scenario-based learning method on the students' academic achievement. *Journal of Baltic Science Education*, *18*(2), 171-183. https://dx.doi.org/10.33225/jbse/19.18.171

Baranger, D. A., Finsaas, M., Goldstein, B., Vize, C., Lynam, D., & Olino, T. (2022, August 4). Tutorial: Power analyses for interaction effects in cross-sectional regressions. https://doi.org/10.31234/osf.io/5ptd7

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*(6), 561-571. https://doi.org/10.1001/archpsyc.1961.01710120031004

Bloom, H. S. (1995). Minimum detectable effects a simple way to report the statistical power of experimental designs. *Evaluation Review*, *19*(5), 547-556. https://doi.org/10.1177/0193841X9501900504

Bloom, H. S. (2006). The core analytics of randomized experiments for social research. Mdrc working papers on research methodology. New York, NY: MDRC. Retrieved from https://files.eric.ed.gov/fulltext/ED493363.pdf

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, *5*(1), 43-82. https://doi.org/10.1080/19345747.2011.578707

Bokai, W. A. N. G., Hongyue, W. A. N. G., Xin, M., & Changyong, F. E. N. G. (2017). Comparisons of superiority, non-inferiority, and equivalence trials. *Shanghai Archives of Psychiatry*, *29*(6), 385. https://doi.org/10.11919%2Fj.issn.1002-0829.217163

Bulus, M. (2021). Sample size determination and optimal design of randomized/non-equivalent pretest-posttest control-group designs. *Adiyaman Univesity Journal of Educational Sciences*, *11*(1), 48-69. https://doi.org/10.17984/adyuebd.941434

Bulus, M. (2022). Minimum detectable effect size computations for cluster-level regression discontinuity: Specifications beyond the linear functional form. *Journal of Research on Education Effectiveness*, *15*(1), 151-177. https://doi.org/10.1080/19345747.2021.1947425

Bulus, M. (2023). Statistical power and sample size calculation tools. R package version 0.3.1. https://cran.r-project.org/package=pwrss

Bulus, M., & Dong, N. (2021a). cosa: Bound constrained optimal sample size allocation. R package version 2.1.0. https://CRAN.R-project.org/package=cosa

Bulus, M., & Dong, N. (2021b). Bound constrained optimization of sample sizes subject to monetary restrictions in planning of multilevel randomized trials and regression discontinuity studies. *The Journal of Treatmental Education*, *89*(2), 379–401. https://doi.org/10.1080/00220973.2019.1636197

Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2021). PowerUpR: Power analysis tools for multilevel randomized treatments. R package version 1.1.0. https://CRAN.R-project.org/package=PowerUpR

Bulus, M., & Koyuncu, I. (2021). Statistical power and precision of treatmental studies originated in the Republic of Turkey from 2010 to 2020: Current practices and some recommendations. *Journal of Participatory Education Research*, *8*(4), 24-43. https://doi.org/10.17275/per.21.77.8.4

Bulus, M., & Sahin, S. G. (2019). Estimation and standardization of variance parameters for planning cluster-randomized trials: A short guide for researchers. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(2), 179-201. https://doi.org/10.21031/epod.530642

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2019). Power calculations for regression-discontinuity designs. *The Stata Journal*, *19*(1), 210-245. https://doi.org/10.1177%2F1536867X19830919

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., ... & De Rosario, H. (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). Routledge.

Committee for Proprietary Medicinal Products (CPMP). (1998). *Notes for guidance on statistical principles for clinical trials*. London (UK): European Medicines Agency (EMEA), 37.

Committee for Proprietary Medicinal Products (CPMP). (2001). Points to consider on switching between superiority and non-inferiority. *British Journal of Clinical Pharmacology*, *52*(3), 223. https://doi.org/10.1046/j.0306-5251.2001.01397-3.x

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 studies. *Journal of Research on Educational Effectiveness*, *6*(1), 24-67. https://doi.org/10.1080/19345747.2012.673143

Dong, N., Spybrook, J., Kelcey, B., & Bulus, M. (2021). Power analyses for moderator effects with (non)random slopes in cluster randomized trials. *Methodology*, *17*(2), 92-110. https://doi.org/10.5964/meth.4003

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments & Computers*, *28*(1), 1-11. https://doi.org/10.3758/BF03203630

Gelman, A. (2019). Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery*, *269*(1), e9-e10.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74-78. https://doi.org/10.1016/j.paid.2016.06.069

Hedberg, E. C. (2017). Introduction to power analysis: two-group studies (Vol. 176). Sage Publications.

Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED509387.pdf

Hisli, N. (1989). Beck depresyon envanterinin üniversite katılımcıleri için geçerliliği, güvenilirliği. *Psikoloji Dergisi*, *7*(23), 3-13. https://toad.halileksi.net/sites/default/files/pdf/beck-depresyon-envanteri-toad.pdf

Kartal, G., Babür, N., & Erçetin, G. (2016). Training for phonological awareness in an orthographically transparent language in two different modalities. *Reading & Writing Quarterly*, *32*(6), 550-579. https://doi.org/10.1080/10573569.2015.1065213

Kelcey, B., Dong, N., Spybrook, J., & Cox, K. (2017a). Statistical power for causally defined ındirect effects in group-randomized trials with ındividual-level mediators. *Journal of Educational and Behavioral Statistics*, *42*(5), 499–530. https://doi.org/10.3102/1076998617695506

Kelcey, B., Dong, N., Spybrook, J., & Shen, Z. (2017b). Experimental power for ındirect effects in group-randomized studies with group-level mediators. *Multivariate Behavioral Research*, *52*(6), 699–719. https://doi.org/10.1080/00273171.2017.1356212

Konstantopoulos, S. (2008a). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, *1*, 265–288. https://doi.org/10.1080/19345740802328216

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness*, *1*, 66–88. https://doi.org/10.1080/19345740701692522

Liu, X. S. (2013). *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. Routledge.

Myors, B., Murphy, K. R., & Wolach, A. (2023). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (5th ed.). Routledge.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide* (7th ed). Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_7.pdf

Otanga, H., Tanhan, A., Musılı, P.M., Arslan, G, & Bulus, M. (2021). Exploring college students' biopsychosocial spiritual wellbeing and problems during COVID-19-19 through a contextual and comprehensive framework. *Journal of Mental Health and Addiction*, *20*, 619-638. https://doi.org/10.1007/s11469-021-00687-9

Ozcan, B., & Bulus, M. (2022). Protective factors associated with academic resilience of adolescents in individualist and collectivist cultures: Evidence from PISA 2018 large scale assessment. *Current Psychology*, *41*, 1740-1756. https://doi.org/10.1007/s12144-022-02944-z

Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. A. (2021). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia Medica*, *31*(1), 27-53. https://doi.org/10.11613/bm.2021.010502

Schochet, P. Z. (2008). *Technical methods report: statistical power for regression discontinuity designs in education evaluations* (NCEE 2008-4026). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://files.eric.ed.gov/fulltext/ED511782.pdf

Schochet, P. Z. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics*, *34*(2), 238-266. Retrieved from http://www.jstor.org/stable/40263528

Şevgin, H. & Çetin, B. (2017). Eğitim araştırmalarında güç analizi ve bir uygulama. *Van Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, *14*(1), 1462-1480. https://dergipark.org.tr/tr/pub/yyuefd/issue/28496/360587

Milli Eğitim Bakanlığı (MEB) (2021). Özel öğrenme güçlüğü olan bireyler: "Aileler için rehber kitapçık". Özel Eğitim ve Rehberlik Hizmetleri Genel Müdürlüğü. https://orgm.meb.gov.tr/meb_iys_dosyalar/2021_02/04102620_OYRENME_GUCLUYU_OLAN_BYREYLER__TR.pdf

Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177%2F2515245920918253

Zhang, Z., & Yuan, K. H. (2018). *Practical statistical power analysis using Webpower and R*. Isdsa Press.