

Package ‘samplesize4surveys’

January 17, 2020

Type Package

Title Sample Size Calculations for Complex Surveys

Version 4.1.1

Date 2020-01-16

Author Hugo Andres Gutierrez Rojas

Maintainer Hugo Andres Gutierrez Rojas <hagutierrezro@gmail.com>

Description Computes the required sample size for estimation of totals, means and proportions under complex sampling designs.

License GPL (>= 2)

Depends R (>= 3.1), TeachingSampling, timeDate, dplyr, magrittr

Suggests knitr

LazyData true

VignetteBuilder knitr

RoxygenNote 7.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2020-01-17 13:20:02 UTC

R topics documented:

b4ddm	2
b4ddp	4
b4dm	5
b4dp	7
b4m	8
b4p	9
b4S2	10
BigLucyT0T1	12
DEFF	13
e4ddm	15
e4ddp	17

e4dm	18
e4dp	19
e4m	21
e4p	22
e4S2	23
ICC	24
ss2s4m	26
ss2s4p	30
ss4ddm	31
ss4ddmH	34
ss4ddp	36
ss4ddpH	39
ss4dm	41
ss4dmH	43
ss4dp	45
ss4dpH	47
ss4HHSm	49
ss4HHSp	51
ss4m	53
ss4mH	55
ss4p	56
ss4pH	58
ss4pLN	59
ss4S2	61
ss4S2H	63
ss4stm	64

Index **67**

b4ddm	<i>Statistical power for a hypothesis testing on a double difference of means.</i>
-------	--

Description

This function computes the power for a (right tail) test of double difference of means

Usage

```
b4ddm(
  N,
  n,
  mu1,
  mu2,
  mu3,
  mu4,
  sigma1,
  sigma2,
```

```

    sigma3,
    sigma4,
    D,
    DEFF = 1,
    conf = 0.95,
    T = 0,
    R = 1,
    plot = FALSE
)

```

Arguments

N	The population size.
n	The sample size.
mu1	The value of the estimated mean of the variable of interest for the first population.
mu2	The value of the estimated mean of the variable of interest for the second population.
mu3	The value of the estimated mean of the variable of interest for the third population.
mu4	The value of the estimated mean of the variable of interest for the fourth population.
sigma1	The value of the estimated variance of the variable of interest for the first population.
sigma2	The value of the estimated mean of a variable of interest for the second population.
sigma3	The value of the estimated variance of the variable of interest for the third population.
sigma4	The value of the estimated mean of a variable of interest for the fourth population.
D	The value of the null effect.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
T	The overlap between waves. By default T = 0.
R	The correlation between waves. By default R = 1.
plot	Optionally plot the power achieved for an specific sample size.

Details

We note that the power is defined as:

$$1 - \Phi\left(Z_{1-\alpha} - \frac{(D - [(\mu_1 - \mu_2) - (\mu_3 - \mu_4)])}{\sqrt{\frac{1}{n}\left(1 - \frac{n}{N}\right)S^2}}\right)$$

where

$$S^2 = DEFF(\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2)$$

Value

The power of the test.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```
b4ddm(N = 100000, n = 400, mu1=50, mu2=55, mu3=50, mu4=55,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12, D = 7)
b4ddm(N = 100000, n = 400, mu1=50, mu2=55, mu3=50, mu4=65,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12, D = 12, plot = TRUE)
b4ddm(N = 100000, n = 4000, mu1=50, mu2=55, mu3=50, mu4=65,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12, D = 11, DEFF = 2, conf = 0.99, plot = TRUE)
```

b4ddp

Statistical power for a hyphotesis testing on a difference of proportions

Description

This function computes the power for a (right tail) test of difference of proportions.

Usage

```
b4ddp(N, n, P1, P2, P3, P4, D, DEFF = 1, conf = 0.95, plot = FALSE)
```

Arguments

N	The population size.
n	The sample size.
P1	The value of the first estimated proportion.
P2	The value of the second estimated proportion.
P3	The value of the third estimated proportion.
P4	The value of the fourth estimated proportion.
D	The value of the null effect.

DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
plot	Optionally plot the power achieved for an specific sample size.

Details

We note that the power is defined as:

$$1 - \Phi\left(Z_{1-\alpha} - \frac{(D - [(P_1 - P_2) - (P_3 - P_4)])}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2 + P_3Q_3 + P_4Q_4)}}\right)$$

Value

The power of the test.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```
b4ddp(N = 10000, n = 400, P1 = 0.5, P2 = 0.5, P3 = 0.5, P4 = 0.5, D = 0.03)
b4ddp(N = 10000, n = 400, P1 = 0.5, P2 = 0.5, P3 = 0.5, P4 = 0.5, D = 0.03, plot = TRUE)
b4ddp(N = 10000, n = 4000, P1 = 0.5, P2 = 0.5, P3 = 0.5, P4 = 0.5,
D = 0.05, DEFF = 2, conf = 0.99, plot = TRUE)
```

b4dm

Statistical power for a hypothesis testing on a difference of means.

Description

This function computes the power for a (right tail) test of difference of means

Usage

```
b4dm(N, n, mu1, mu2, sigma1, sigma2, D, DEFF = 1, conf = 0.95, plot = FALSE)
```

Arguments

N	The population size.
n	The sample size.
mu1	The value of the estimated mean of the variable of interes for the first population.
mu2	The value of the estimated mean of the variable of interes for the second population.
sigma1	The value of the estimated variance of the variable of interes for the first population.
sigma2	The value of the estimated mean of a variable of interes for the second population.
D	The value of the null effect.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
plot	Optionally plot the power achieved for an specific sample size.

Details

We note that the power is defined as:

$$1 - \Phi\left(Z_{1-\alpha} - \frac{(D - (\mu_1 - \mu_2))}{\sqrt{\frac{1}{n}\left(1 - \frac{n}{N}\right)S^2}}\right)$$

where

$$S^2 = DEFF(\sigma_1^2 + \sigma_2^2)$$

Value

The power of the test.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```

b4dm(N = 100000, n = 400, mu1 = 5, mu2 = 5, sigma1 = 10, sigma2 = 15, D = 5)
b4dm(N = 100000, n = 400, mu1 = 5, mu2 = 5, sigma1 = 10, sigma2 = 15, D = 0.03, plot = TRUE)
b4dm(N = 100000, n = 4000, mu1 = 5, mu2 = 5, sigma1 = 10, sigma2 = 15,
D = 0.05, DEFF = 2, conf = 0.99, plot = TRUE)

```

b4dp

*Statistical power for a hypothesis testing on a difference of proportions***Description**

This function computes the power for a (right tail) test of difference of proportions.

Usage

```
b4dp(N, n, P1, P2, D, DEFF = 1, conf = 0.95, plot = FALSE)
```

Arguments

N	The population size.
n	The sample size.
P1	The value of the first estimated proportion.
P2	The value of the second estimated proportion.
D	The value of the null effect.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
plot	Optionally plot the power achieved for an specific sample size.

Details

We note that the power is defined as:

$$1 - \Phi\left(Z_{1-\alpha} - \frac{(D - (P_1 - P_2))}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2)}}\right)$$

Value

The power of the test.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```
b4dp(N = 100000, n = 400, P1 = 0.5, P2 = 0.5, D = 0.03)
b4dp(N = 100000, n = 400, P1 = 0.5, P2 = 0.5, D = 0.03, plot = TRUE)
b4dp(N = 100000, n = 4000, P1 = 0.5, P2 = 0.5, D = 0.05, DEFF = 2, conf = 0.99, plot = TRUE)
```

b4m

Statistical power for a hypothesis testing on a single mean

Description

This function computes the power for a (right tail) test of means.

Usage

```
b4m(N, n, mu, sigma, D, DEFF = 1, conf = 0.95, plot = FALSE)
```

Arguments

N	The population size.
n	The sample size.
mu	The value of the estimated mean of the variable of interest.
sigma	The value of the standard deviation of the variable of interest.
D	The value of the null effect. Note that D must be strictly greater than mu.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
plot	Optionally plot the power achieved for an specific sample size.

Details

We note that the power is defined as:

$$1 - \Phi\left(Z_{1-\alpha} - \frac{(D - \mu)}{\sqrt{\frac{1}{n}\left(1 - \frac{n}{N}\right)S^2}}\right)$$

where

$$S^2 = DEFF\sigma^2$$

Value

The power of the test.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```
b4m(N = 100000, n = 400, mu = 3, sigma = 1, D = 3.1)
b4m(N = 100000, n = 400, mu = 5, sigma = 10, D = 7, plot = TRUE)
b4m(N = 100000, n = 400, mu = 50, sigma = 100, D = 100, DEFF = 3.4, conf = 0.99, plot = TRUE)
```

b4p

Statistical power for a hypothesis testing on a single proportion

Description

This function computes the power for a (right tail) test of proportions.

Usage

```
b4p(N, n, P, D, DEFF = 1, conf = 0.95, plot = FALSE)
```

Arguments

N	The population size.
n	The sample size.
P	The value of the first estimated proportion.
D	The value of the null effect. Note that D must be strictly greater than P.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
plot	Optionally plot the power achieved for an specific sample size.

Details

We note that the power is defined as:

$$1 - \Phi\left(Z_{1-\alpha} - \frac{(D - P)}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P(1 - P))}}\right)$$

Value

The power of the test.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```
b4p(N = 100000, n = 400, P = 0.5, D = 0.55)
b4p(N = 100000, n = 400, P = 0.5, D = 0.9, plot = TRUE)
b4p(N = 100000, n = 4000, P = 0.5, D = 0.55, DEFF = 2, conf = 0.99, plot = TRUE)
```

b4S2

Statistical power for a hypothesis testing on a single variance

Description

This function computes the power for a (right tail) test of variance

Usage

```
b4S2(N, n, S2, S20, K = 0, DEFF = 1, conf = 0.95, power = 0.8, plot = FALSE)
```

Arguments

N	The population size.
n	The sample size.
S2	The value of the first estimated proportion.
S20	The value of the null effect. Note that S2 must be strictly smaller than S2.
K	The excess kurtosis of the variable in the population.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
power	The statistical power. By default power = 0.80.
plot	Optionally plot the power achieved for an specific sample size.

Details

We note that the power is defined as:

$$1 - \Phi\left(Z_{1-\alpha} - \frac{(D - P)}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P(1 - P))}}\right)$$

Value

The power of the test.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```
b4S2(N = 100000, n = 400, S2 = 120, S20 = 100, K = 0, DEFF = 1)
b4S2(N = 100000, n = 400, S2 = 120, S20 = 100, K = 2, DEFF = 1)
b4S2(N = 100000, n = 400, S2 = 120, S20 = 100, K = 2, DEFF = 2.5, plot = TRUE)
```

BigLucyT0T1

*Some Business Population Database for two periods of time***Description**

This data set corresponds to a random sample of BigLucy. It contains some financial variables of 85296 industrial companies of a city in a particular fiscal year.

Usage

BigLucyT0T1

Format

ID The identifier of the company. It correspond to an alphanumeric sequence (two letters and three digits)

Ubication The address of the principal office of the company in the city

Level The industrial companies are discriminated according to the Taxes declared. There are small, medium and big companies

Zone The city is divided by geographical zones. A company is classified in a particular zone according to its address

Income The total amount of a company's earnings (or profit) in the previous fiscal year. It is calculated by taking revenues and adjusting for the cost of doing business

Employees The total number of persons working for the company in the previous fiscal year

Taxes The total amount of a company's income Tax

SPAM Indicates if the company uses the Internet and WEBmail options in order to make self-propaganda.

Segments The cartographic divisions.

Outgoing Expenses per year.

Years Age of the company.

ISO Indicates whether the company is quality-certified.

ISOYears Indicates the time company has been certified.

CountyP Indicates whether the county is participating in the intervention. That is if the county contains companies that have been certified by ISO

Time Refers to the time of observation.

Author(s)

Hugo Andres Gutierrez Rojas <hugogutierrez@usantotomas.edu.co>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Editorial Universidad Santo Tomas.

Examples

```

data(Lucy)
attach(Lucy)
# The variables of interest are: Income, Employees and Taxes
# This information is stored in a data frame called estima
estima <- data.frame(Income, Employees, Taxes)
# The population totals
colSums(estima)
# Some parameters of interest
table(SPAM,Level)
xtabs(Income ~ Level+SPAM)
# Correlations among characteristics of interest
cor(estima)
# Some useful histograms
hist(Income)
hist(Taxes)
hist(Employees)
# Some useful plots
boxplot(Income ~ Level)
barplot(table(Level))
pie(table(SPAM))

```

DEFF

*Estimated sample Effects of Design (DEFF)***Description**

This function returns the estimated design effects for a set of inclusion probabilities and the variables of interest.

Usage

```
DEFF(y, pik)
```

Arguments

y	Vector, matrix or data frame containing the recollected information of the variables of interest for every unit in the selected sample.
pik	Vector of inclusion probabilities for each unit in the selected sample.

Details

The design effect is somehow defined to be the ratio between the variance of a complex design and the variance of a simple design. When the design is stratified and the allocation is proportional, this measure reduces to

$$DEFF_{Kish} = 1 + CV(w)$$

where w is the set of weights (defined as the inverse of the inclusion probabilities) along the sample, and CV refers to the classical coefficient of variation. Although this measure is #' motivated by a

stratified sampling design, it is commonly applied to any kind of survey where sampling weight are unequal. On the other hand, the Spencer's DEFF is motivated by the idea that a set of weights may be efficient even when they vary, and is defined by:

$$DEFF_{Spencer} = (1 - R^2) * DEFF_{Kish} + \frac{\hat{a}^2}{\hat{\sigma}_y^2} * (DEFF_{Kish} - 1)$$

where

$$\hat{\sigma}_y^2 = \frac{\sum_s w_k (y_k - \bar{y}_w)^2}{\sum_s w_k}$$

and \hat{a} is the estimation of the intercept in the following model

$$y_k = a + b * p_k + e_k$$

with $p_k = \pi_k/n$ is a standardized sampling weight. Finally, R^2 is the R-squared of this model.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Editorial Universidad Santo Tomás. Valliant, R, et. al. (2013), *Practical tools for Design and Weighting Survey Samples*. Springer

Examples

```
#####
# Example with BigLucy data #
#####
data(BigLucy)
attach(BigLucy)

# The sample size
n <- 400
res <- S.piPS(n, Income)
sam <- res[,1]
# The information about the units in the sample is stored in an object called data
data <- BigLucy[sam,]
attach(data)
names(data)
# pik.s is the inclusion probability of every single unit in the selected sample
pik <- res[,2]
# The variables of interest are: Income, Employees and Taxes
# This information is stored in a data frame called estima
estima <- data.frame(Income, Employees, Taxes)
E.piPS(estima,pik)
DEFF(estima,pik)
```

e4ddm

Statistical errors for the estimation of a double difference of means

Description

This function computes the coefficient of variation and the standard error when estimating a double difference of means under a complex sample design.

Usage

```
e4ddm(  
  N,  
  n,  
  mu1,  
  mu2,  
  mu3,  
  mu4,  
  sigma1,  
  sigma2,  
  sigma3,  
  sigma4,  
  DEFF = 1,  
  conf = 0.95,  
  T = 0,  
  R = 1,  
  plot = FALSE  
)
```

Arguments

N	The population size.
n	The sample size.
mu1	The value of the estimated mean of the variable of interest for the first population.
mu2	The value of the estimated mean of the variable of interest for the second population.
mu3	The value of the estimated mean of the variable of interest for the third population.
mu4	The value of the estimated mean of the variable of interest for the fourth population.
sigma1	The value of the estimated variance of the variable of interest for the first population.
sigma2	The value of the estimated mean of a variable of interest for the second population.
sigma3	The value of the estimated variance of the variable of interest for the third population.

sigma4	The value of the estimated mean of a variable of interest for the fourth population.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
T	The overlap between waves. By default T = 0.
R	The correlation between waves. By default R = 1.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

We note that the coefficient of variation is defined as:

$$cve = \frac{\sqrt{Var((\bar{y}_1 - \bar{y}_2) - (\bar{y}_3 - \bar{y}_4))}}{(\bar{y}_1 - \bar{y}_2) - (\bar{y}_3 - \bar{y}_4)}$$

Also, note that the margin of error is defined as:

$$\varepsilon = z_{1-\frac{\alpha}{2}} \sqrt{Var((\bar{y}_1 - \bar{y}_2) - (\bar{y}_3 - \bar{y}_4))}$$

Value

The coefficient of variation and the margin of error for a predefined sample size.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Editorial Universidad Santo Tomás

See Also

[ss4p](#)

Examples

```
e4ddm(N=10000, n=400, mu1=50, mu2=55, mu3=50, mu4=65,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12)
e4ddm(N=10000, n=400, mu1=50, mu2=55, mu3=50, mu4=65,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12, plot=TRUE)
e4ddm(N=10000, n=400, mu1=50, mu2=55, mu3=50, mu4=65,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12, DEFF=3.45, conf=0.99, plot=TRUE)
```

e4ddp	<i>Statistical errors for the estimation of a double difference of proportions</i>
-------	--

Description

This function computes the coefficient of variation and the standard error when estimating a double difference of proportions under a complex sample design.

Usage

```
e4ddp(N, n, P1, P2, P3, P4, DEFF = 1, conf = 0.95, plot = FALSE)
```

Arguments

N	The population size.
n	The sample size.
P1	The value of the first estimated proportion.
P2	The value of the second estimated proportion.
P3	The value of the third estimated proportion.
P4	The value of the fourth estimated proportion.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

We note that the margin of error is defined as:

$$cve = \frac{\sqrt{\text{Var}((\hat{P}_1 - \hat{P}_2) - (\hat{P}_3 - \hat{P}_4))}}{(\hat{P}_1 - \hat{P}_2) - (\hat{P}_3 - \hat{P}_4)}$$

Also, note that the margin of error is defined as:

$$\varepsilon = z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}((\hat{P}_1 - \hat{P}_2) - (\hat{P}_3 - \hat{P}_4))}$$

Value

The coefficient of variation and the margin of error for a predefined sample size.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```
e4ddp(N=10000, n=400, P1=0.5, P2=0.6, P3=0.5, P4=0.7)
e4ddp(N=10000, n=400, P1=0.5, P2=0.6, P3=0.5, P4=0.7, plot=TRUE)
e4ddp(N=10000, n=400, P1=0.5, P2=0.6, P3=0.5, P4=0.7, DEFF=3.45, conf=0.99, plot=TRUE)
```

e4dm

Statistical errors for the estimation of a difference of means

Description

This function computes the coefficient of variation and the standard error when estimating a difference of means under a complex sample design.

Usage

```
e4dm(N, n, mu1, mu2, sigma1, sigma2, DEFF = 1, conf = 0.95, plot = FALSE)
```

Arguments

N	The population size.
n	The sample size.
mu1	The value of the estimated mean of the variable of interest for the first population.
mu2	The value of the estimated mean of the variable of interest for the second population.
sigma1	The value of the estimated variance of the variable of interest for the first population.
sigma2	The value of the estimated mean of a variable of interest for the second population.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

We note that the coefficient of variation is defined as:

$$cve = \frac{\sqrt{Var(\bar{y}_1 - \bar{y}_2)}}{\bar{y}_1 - \bar{y}_2}$$

Also, note that the margin of error is defined as:

$$\varepsilon = z_{1-\frac{\alpha}{2}} \sqrt{Var(\bar{y}_1 - \bar{y}_2)}$$

Value

The coefficient of variation and the margin of error for a predefined sample size.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```
e4dm(N=10000, n=400, mu1 = 100, mu2 = 12, sigma1 = 10, sigma2=8)
e4dm(N=10000, n=400, mu1 = 100, mu2 = 12, sigma1 = 10, sigma2=8, plot=TRUE)
e4dm(N=10000, n=400, mu1 = 100, mu2 = 12, sigma1 = 10, sigma2=8, DEFF=3.45, conf=0.99, plot=TRUE)
```

e4dp

Statistical errors for the estimation of a difference of proportions

Description

This function computes the coefficient of variation and the standard error when estimating a difference of proportions under a complex sample design.

Usage

```
e4dp(N, n, P1, P2, DEFF = 1, T = 0, R = 1, conf = 0.95, plot = FALSE)
```

Arguments

N	The population size.
n	The sample size.
P1	The value of the first estimated proportion.
P2	The value of the second estimated proportion.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
T	The overlap between waves. By default T = 0.
R	The correlation between waves. By default R = 1.
conf	The statistical confidence. By default conf = 0.95.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

We note that the margin of error is defined as:

$$cve = \frac{\sqrt{Var(\hat{P}_1 - \hat{P}_2)}}{\hat{P}_1 - \hat{P}_2}$$

Also, note that the margin of error is defined as:

$$\varepsilon = z_{1-\frac{\alpha}{2}} \sqrt{Var(\hat{P}_1 - \hat{P}_2)}$$

Value

The coefficient of variation and the margin of error for a predefined sample size.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro@gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```
e4dp(N=10000, n=400, P1=0.5, P2=0.6)
e4dp(N=10000, n=400, P1=0.5, P2=0.6, plot=TRUE)
e4dp(N=10000, n=400, P1=0.5, P2=0.6, DEFF=3.45, conf=0.99, plot=TRUE)
e4dp(N=10000, n=400, P1=0.5, P2=0.6, T=0.5, R=0.5, DEFF=3.45, conf=0.99, plot=TRUE)
```

Description

This function computes the coefficient of variation and the standard error when estimating a single mean under a complex sample design.

Usage

```
e4m(N, n, mu, sigma, DEFF = 1, conf = 0.95, plot = FALSE)
```

Arguments

N	The population size.
n	The sample size.
mu	The value of the estimated mean of the variable of interest.
sigma	The value of the standard deviation of the variable of interest.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

We note that the coefficient of variation is defined as:

$$cve = \frac{\sqrt{Var(\bar{y}_S)}}{\bar{y}_S}$$

Also, note that the margin of error is defined as:

$$\varepsilon = z_{1-\frac{\alpha}{2}} \sqrt{Var(\bar{y}_S)}$$

Value

The coefficient of variation and the margin of error for a predefined sample size.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also[ss4p](#)**Examples**

```
e4m(N=10000, n=400, mu = 10, sigma = 10)
e4m(N=10000, n=400, mu = 10, sigma = 10, plot=TRUE)
e4m(N=10000, n=400, mu = 10, sigma = 10, DEFF=3.45, conf=0.99, plot=TRUE)
```

e4p

*Statistical errors for the estimation of a single proportion***Description**

This function computes the coefficient of variation and the standard error when estimating a single proportion under a sample design.

Usage

```
e4p(N, n, P, DEFF = 1, conf = 0.95, plot = FALSE)
```

Arguments

N	The population size.
n	The sample size.
P	The value of the estimated proportion.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

We note that the coefficient of variation is defined as:

$$cve = \frac{\sqrt{Var(\hat{p})}}{\hat{p}}$$

Also, note that the margin of error is defined as:

$$\varepsilon = z_{1-\frac{\alpha}{2}} \sqrt{Var(\hat{p})}$$

Value

The coefficient of variation, the margin of error and the relative margin of error for a predefined sample size.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```
e4p(N=10000, n=400, P=0.5)
e4p(N=10000, n=400, P=0.5, plot=TRUE)
e4p(N=10000, n=400, P=0.01, DEFF=3.45, conf=0.99, plot=TRUE)
```

e4S2

Statistical errors for the estimation of a single variance

Description

This function computes the coefficient of variation and the margin of error when estimating a single variance under a sample design.

Usage

```
e4S2(N, n, K = 0, DEFF = 1, conf = 0.95, plot = FALSE)
```

Arguments

N	The population size.
n	The sample size.
K	The excess kurtosis of the variable in the population.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

We note that the coefficient of variation is defined as:

$$cve = \frac{\sqrt{Var(\hat{S}^2)}}{\hat{S}^2}$$

Also, note that the margin of error is defined as:

$$\varepsilon = z_{1-\frac{\alpha}{2}} \sqrt{Var(\hat{S}^2)}$$

Value

The coefficient of variation and the margin of error for a predefined sample size.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```
e4S2(N=10000, n=400, K = 0)
e4S2(N=10000, n=400, K = 1, DEFF = 2, conf = 0.99)
e4S2(N=10000, n=400, K = 2, DEFF = 2, conf = 0.99, plot=TRUE)
```

 ICC

Intraclass Correlation Coefficient

Description

This function computes the intraclass correlation coefficient.

Usage

```
ICC(y, c1)
```

Arguments

y The variable of interest.
c1 The variable indicating the membership of each element to a specific cluster.

Details

The intraclass correlation coefficient is defined as:

$$\rho = 1 - \frac{m}{m-1} \frac{WSS}{TSS}$$

Where m is the average sample size of units selected inside each sampled cluster.

Value

The total sum of squares (TSS), the between sum of squares (BSS), the within sum of squares (WSS) and the intraclass correlation coefficient.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Editorial Universidad Santo Tomás

See Also

[ss4p](#)

Examples

```
#####
# Almost same mean in each cluster      #
#                                       #
# - Heterogeneity within clusters       #
# - Homogeneity between clusters        #
#####

# Population size
N <- 100000
# Number of clusters in the population
NI <- 1000
# Number of elements per cluster
N/NI

# The variable of interest
y <- c(1:N)
# The clustering factor
cl <- rep(1:NI, length.out=N)

table(cl)
tapply(y, cl, FUN=mean)
boxplot(y~cl)
rho = ICC(y,cl)$ICC
```

```

rho

#####
# Very different means per cluster      #
#                                       #
# - Heterogeneity between clusters     #
# - Homogeneity within clusters        #
#####

# Population size
N <- 100000
# Number of clusters in the population
NI <- 1000
# Number of elements per cluster
N/NI

# The variable of interest
y <- c(1:N)
# The clustering factor
cl <- kronecker(c(1:NI),rep(1,N/NI))

table(cl)
tapply(y, cl, FUN=mean)
boxplot(y~cl)
rho = ICC(y,cl)$ICC
rho

#####
# Example 1 with Lucy data #
#####

data(Lucy)
attach(Lucy)
N <- nrow(Lucy)
y <- Income
cl <- Zone
ICC(y,cl)

#####
# Example 2 with Lucy data #
#####

data(Lucy)
attach(Lucy)
N <- nrow(Lucy)
y <- as.double(SPAM)
cl <- Zone
ICC(y,cl)

```

ss2s4m

*Sample Sizes in Two-Stage sampling Designs for Estimating Single Means***Description**

This function computes a grid of possible sample sizes for estimating single means under two-stage sampling designs.

Usage

```
ss2s4m(N, mu, sigma, conf = 0.95, delta = 0.03, M, to = 20, rho)
```

Arguments

N	The population size.
mu	The value of the estimated mean of a variable of interest.
sigma	The value of the estimated standard deviation of a variable of interest.
conf	The statistical confidence. By default conf = 0.95. By default conf = 0.95.
delta	The maximum relative margin of error that can be allowed for the estimation.
M	Number of clusters in the population.
to	(integer) maximum number of final units to be selected per cluster. By default to = 20.
rho	The Intraclass Correlation Coefficient.

Details

In two-stage (2S) sampling, the design effect is defined by

$$DEFF = 1 + (m - 1)\rho$$

Where ρ is defined as the intraclass correlation coefficient, m is the average sample size of units selected inside each cluster. The relationship of the full sample size of the two stage design (2S) with the simple random sample (SI) design is given by

$$n_{2S} = n_{SI} * DEFF$$

Value

This function returns a grid of possible sample sizes. The first column represent the design effect, the second column is the number of clusters to be selected, the third column is the number of units to be selected inside the clusters, and finally, the last column indicates the full sample size induced by this particular strategy.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ICC](#)

Examples

```
ss2s4m(N=100000, mu=10, sigma=2, conf=0.95, delta=0.03, M=50, rho=0.01)
ss2s4m(N=100000, mu=10, sigma=2, conf=0.95, delta=0.03, M=50, to=40, rho=0.1)
ss2s4m(N=100000, mu=10, sigma=2, conf=0.95, delta=0.03, M=50, to=40, rho=0.2)
ss2s4m(N=100000, mu=10, sigma=2, conf=0.95, delta=0.05, M=50, to=40, rho=0.3)
```

```
#####
# Almost same mean in each cluster      #
#                                         #
# - Heterogeneity within clusters       #
# - Homogeneity between clusters        #
#                                         #
# Decision rule:                         #
# * Select a lot of units per cluster    #
# * Select a few of clusters             #
#####
```

```
# Population size
N <- 100000
# Number of clusters in the population
M <- 1000
# Number of elements per cluster
N/M
```

```
# The variable of interest
y <- c(1:N)
# The clustering factor
cl <- rep(1:M, length.out=N)
```

```
rho = ICC(y,cl)$ICC
rho
```

```
ss2s4m(N, mu=mean(y), sigma=sd(y), conf=0.95, delta=0.03, M=M, rho=rho)
```

```
#####
# Very different means per cluster      #
#                                         #
# - Heterogeneity between clusters       #
# - Homogeneity within clusters          #
#                                         #
# Decision rule:                         #
#####
```

```

# * Select a few of units per cluster #
# * Select a lot of clusters #
#####

# Population size
N <- 1000000
# Number of clusters in the population
M <- 1000
# Number of elements per cluster
N/M

# The variable of interest
y <- c(1:N)
# The clustering factor
cl <- kronecker(c(1:M),rep(1,N/M))

rho = ICC(y,cl)$ICC
rho

ss2s4m(N, mu=mean(y), sigma=sd(y), conf=0.95, delta=0.03, M=M, rho=rho)

#####
# Example with Lucy data #
#####

data(BigLucy)
attach(BigLucy)
N <- nrow(BigLucy)
P <- prop.table(table(SPAM))[1]
y <- Income
cl <- Segments

rho <- ICC(y,cl)$ICC
M <- length(levels(Segments))

ss2s4m(N, mu=mean(y), sigma=sd(y), conf=0.95, delta=0.03, M=M, rho=rho)

#####
# Example with Lucy data #
#####

data(BigLucy)
attach(BigLucy)
N <- nrow(BigLucy)
P <- prop.table(table(SPAM))[1]
y <- Years
cl <- Segments

rho <- ICC(y,cl)$ICC
M <- length(levels(Segments))

ss2s4m(N, mu=mean(y), sigma=sd(y), conf=0.95, delta=0.03, M=M, rho=rho)

```

 ss2s4p

Sample Sizes in Two-Stage sampling Designs for Estimating Single Proportions

Description

This function computes a grid of possible sample sizes for estimating single proportions under two-stage sampling designs.

Usage

```
ss2s4p(N, P, conf = 0.95, delta = 0.03, M, to = 20, rho)
```

Arguments

N	The population size.
P	The value of the estimated proportion.
conf	The statistical confidence. By default conf = 0.95.
delta	The maximum margin of error that can be allowed for the estimation.
M	Number of clusters in the population.
to	(integer) maximum number of final units to be selected per cluster. By default to = 20.
rho	The Intraclass Correlation Coefficient.

Details

In two-stage (2S) sampling, the design effect is defined by

$$DEFF = 1 + (\bar{m} - 1)\rho$$

Where ρ is defined as the intraclass correlation coefficient, \bar{m} is the average sample size of units selected inside each cluster. The relationship of the full sample size of the two stage design (2S) with the simple random sample (SI) design is given by

$$n_{2S} = n_{SI} * DEFF$$

Value

This function returns a grid of possible sample sizes. The first column represent the design effect, the second column is the number of clusters to be selected, the third column is the number of units to be selected inside the clusters, and finally, the last column indicates the full sample size induced by this particular strategy.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Editorial Universidad Santo Tomas

See Also

[ICC](#)

Examples

```
ss2s4p(N=100000, P=0.5, delta=0.05, M=50, rho=0.01)
ss2s4p(N=100000, P=0.5, delta=0.05, M=500, to=40, rho=0.1)
ss2s4p(N=100000, P=0.5, delta=0.03, M=1000, to=100, rho=0.2)

#####
# Example 2 with Lucy data #
#####

data(BigLucy)
attach(BigLucy)
N <- nrow(BigLucy)
P <- prop.table(table(SPAM))[1]
y <- Domains(SPAM)[, 1]
cl <- Segments

rho <- ICC(y,cl)$ICC
M <- length(levels(Segments))
ss2s4p(N, P, conf=0.95, delta = 0.03, M=M, to=30, rho=rho)
```

ss4ddm

The required sample size for estimating a double difference of means

Description

This function returns the minimum sample size required for estimating a double difference of means subject to predefined errors.

Usage

```
ss4ddm(
  N,
  mu1,
  mu2,
  mu3,
  mu4,
  sigma1,
  sigma2,
```

```

sigma3,
sigma4,
DEFF = 1,
conf = 0.95,
cve = 0.05,
rme = 0.03,
T = 0,
R = 1,
plot = FALSE
)

```

Arguments

N	The maximum population size between the groups (strata) that we want to compare.
mu1	The value of the estimated mean of the variable of interest for the first population.
mu2	The value of the estimated mean of the variable of interest for the second population.
mu3	The value of the estimated mean of the variable of interest for the third population.
mu4	The value of the estimated mean of the variable of interest for the fourth population.
sigma1	The value of the estimated variance of the variable of interest for the first population.
sigma2	The value of the estimated mean of a variable of interest for the second population.
sigma3	The value of the estimated variance of the variable of interest for the third population.
sigma4	The value of the estimated mean of a variable of interest for the fourth population.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95. By default conf = 0.95.
cve	The maximum coefficient of variation that can be allowed for the estimation.
rme	The maximum relative margin of error that can be allowed for the estimation.
T	The overlap between waves. By default T = 0.
R	The correlation between waves. By default R = 1.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

Note that the minimum sample size to achieve a relative margin of error ε is defined by:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Where

$$n_0 = \frac{z_{1-\frac{\alpha}{2}}^2 S^2}{\varepsilon^2 \mu^2}$$

and $S^2 = (\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2) * (1 - (T * R)) * DEFF$ Also note that the minimum sample size to achieve a coefficient of variation *cve* is defined by:

$$n = \frac{S^2}{|(\bar{y}_1 - \bar{y}_2) - (\bar{y}_3 - \bar{y}_4)|^2 cve^2 + \frac{S^2}{N}}$$

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[e4p](#)

Examples

```
ss4ddm(N=100000, mu1=50, mu2=55, mu3=50, mu4=65,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12, cve=0.05, rme=0.03)
ss4ddm(N=100000, mu1=50, mu2=55, mu3=50, mu4=65,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12, cve=0.05, rme=0.03, plot=TRUE)
ss4ddm(N=100000, mu1=50, mu2=55, mu3=50, mu4=65,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12, DEFF=3.45, conf=0.99, cve=0.03,
rme=0.03, plot=TRUE)
```

```
#####
# Example with BigLucy data #
#####
data(BigLucyT0T1)
attach(BigLucyT0T1)
```

```
BigLucyT0 <- BigLucyT0T1[Time == 0,]
BigLucyT1 <- BigLucyT0T1[Time == 1,]
N1 <- table(BigLucyT0$ISO)[1]
N2 <- table(BigLucyT0$ISO)[2]
N <- max(N1,N2)
```

```
BigLucyT0.yes <- subset(BigLucyT0, ISO == "yes")
BigLucyT0.no <- subset(BigLucyT0, ISO == "no")
BigLucyT1.yes <- subset(BigLucyT1, ISO == "yes")
BigLucyT1.no <- subset(BigLucyT1, ISO == "no")
mu1 <- mean(BigLucyT0.yes$Income)
mu2 <- mean(BigLucyT0.no$Income)
mu3 <- mean(BigLucyT1.yes$Income)
```

```

mu4 <- mean(BigLucyT1.no$Income)
sigma1 <- sd(BigLucyT0.yes$Income)
sigma2 <- sd(BigLucyT0.no$Income)
sigma3 <- sd(BigLucyT1.yes$Income)
sigma4 <- sd(BigLucyT1.no$Income)

# The minimum sample size for simple random sampling
ss4ddm(N, mu1, mu2, mu3, mu4, sigma1, sigma2, sigma3, sigma4,
DEFF=1, conf=0.95, cve=0.001, rme=0.001, plot=TRUE)
# The minimum sample size for a complex sampling design
ss4ddm(N, mu1, mu2, mu3, mu4, sigma1, sigma2, sigma3, sigma4,
DEFF=3.45, conf=0.99, cve=0.03, rme=0.03, plot=TRUE)

```

ss4ddmH

The required sample size for testing a null hypothesis for a double difference of proportions

Description

This function returns the minimum sample size required for testing a null hypothesis regarding a double difference of proportions.

Usage

```

ss4ddmH(
  N,
  mu1,
  mu2,
  mu3,
  mu4,
  sigma1,
  sigma2,
  sigma3,
  sigma4,
  D,
  DEFF = 1,
  conf = 0.95,
  power = 0.8,
  T = 0,
  R = 1,
  plot = FALSE
)

```

Arguments

N	The maximum population size between the groups (strata) that we want to compare.
mu1	The value of the estimated mean of the variable of interest for the first population.

mu2	The value of the estimated mean of the variable of interes for the second population.
mu3	The value of the estimated mean of the variable of interes for the third population.
mu4	The value of the estimated mean of the variable of interes for the fourth population.
sigma1	The value of the estimated variance of the variable of interes for the first population.
sigma2	The value of the estimated mean of a variable of interes for the second population.
sigma3	The value of the estimated variance of the variable of interes for the third population.
sigma4	The value of the estimated mean of a variable of interes for the fourth population.
D	The minimun effect to test.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
power	The statistical power. By default power = 0.80.
T	The overlap between waves. By default T = 0.
R	The correlation between waves. By default R = 1.
plot	Optionally plot the effect against the sample size.

Details

We assume that it is of interest to test the following set of hyphotesis:

$$H_0 : (\mu_1 - \mu_2) - (\mu_3 - \mu_4) = 0 \quad vs. \quad H_a : (\mu_1 - \mu_2) - (\mu_3 - \mu_4) = D \neq 0$$

Note that the minimun sample size, restricted to the predefined power β and confidence $1 - \alpha$, is defined by:

$$n = \frac{S^2}{\frac{D^2}{(z_{1-\alpha} + z_\beta)^2} + \frac{S^2}{N}}$$

$$\text{where } S^2 = (\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2) * (1 - (T * R)) * DEFF$$

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4pH](#)

Examples

```
ss4ddmH(N = 100000, mu1=50, mu2=55, mu3=50, mu4=65,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12, D=3)
ss4ddmH(N = 100000, mu1=50, mu2=55, mu3=50, mu4=65,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12, D=1, plot=TRUE)
ss4ddmH(N = 100000, mu1=50, mu2=55, mu3=50, mu4=65,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12, D=0.5, DEFF = 2, plot=TRUE)
ss4ddmH(N = 100000, mu1=50, mu2=55, mu3=50, mu4=65,
sigma1 = 10, sigma2 = 12, sigma3 = 10, sigma4 = 12, D=0.5, DEFF = 2, conf = 0.99,
power = 0.9, plot=TRUE)
```

```
#####
# Example with BigLucy data #
#####
data(BigLucyT0T1)
attach(BigLucyT0T1)

BigLucyT0 <- BigLucyT0T1[Time == 0,]
BigLucyT1 <- BigLucyT0T1[Time == 1,]
N1 <- table(BigLucyT0$ISO)[1]
N2 <- table(BigLucyT0$ISO)[2]
N <- max(N1,N2)

BigLucyT0.yes <- subset(BigLucyT0, ISO == 'yes')
BigLucyT0.no <- subset(BigLucyT0, ISO == 'no')
BigLucyT1.yes <- subset(BigLucyT1, ISO == 'yes')
BigLucyT1.no <- subset(BigLucyT1, ISO == 'no')
mu1 <- mean(BigLucyT0.yes$Income)
mu2 <- mean(BigLucyT0.no$Income)
mu3 <- mean(BigLucyT1.yes$Income)
mu4 <- mean(BigLucyT1.no$Income)
sigma1 <- sd(BigLucyT0.yes$Income)
sigma2 <- sd(BigLucyT0.no$Income)
sigma3 <- sd(BigLucyT1.yes$Income)
sigma4 <- sd(BigLucyT1.no$Income)

# The minimum sample size for testing
# H_0: (mu_1 - mu_2) - (mu_3 - mu_4) = 0 vs.
# H_a: (mu_1 - mu_2) - (mu_3 - mu_4) = D = 3

ss4ddmH(N, mu1, mu2, mu3, mu4, sigma1, sigma2, sigma3, sigma4,
D = 3, conf = 0.99, power = 0.9, DEFF = 3.45, plot=TRUE)
```

ss4ddp

The required sample size for estimating a double difference of proportions

Description

This function returns the minimum sample size required for estimating a double difference of proportion subjecto to predefined errors.

Usage

```

ss4ddp(
  N,
  P1,
  P2,
  P3,
  P4,
  DEFF = 1,
  conf = 0.95,
  cve = 0.05,
  me = 0.03,
  T = 0,
  R = 1,
  plot = FALSE
)

```

Arguments

N	The population size.
P1	The value of the first estimated proportion at first wave.
P2	The value of the second estimated proportion at first wave.
P3	The value of the first estimated proportion at second wave.
P4	The value of the second estimated proportion at second wave.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95. By default conf = 0.95.
cve	The maximum coefficient of variation that can be allowed for the estimation.
me	The maximum margin of error that can be allowed for the estimation.
T	The overlap between waves. By default T = 0.
R	The correlation between waves. By default R = 1.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

Note that the minimum sample size (for each group at each wave) to achieve a particular margin of error ε is defined by:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Where

$$n_0 = \frac{z_{1-\frac{\alpha}{2}}^2 S^2}{\varepsilon^2}$$

and

$$S^2 = (P1 * Q1 + P2 * Q2 + P3 * Q3 + P4 * Q4) * (1 - (T * R)) * DEFF$$

Also note that the minimum sample size to achieve a particular coefficient of variation *cve* is defined by:

$$n = \frac{S^2}{(ddp)^2 cve^2 + \frac{S^2}{N}}$$

And *ddp* is the expected estimate of the double difference of proportions.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4dp](#)

Examples

```
ss4ddp(N=100000, P1=0.05, P2=0.55, P3= 0.5, P4= 0.6, cve=0.05, me=0.03)
ss4ddp(N=100000, P1=0.05, P2=0.55, P3= 0.5, P4= 0.6, cve=0.05, me=0.03, plot=TRUE)
ss4ddp(N=100000, P1=0.05, P2=0.55, P3= 0.5, P4= 0.6, DEFF=3.45, conf=0.99,
cve=0.03, me=0.03, plot=TRUE)
ss4ddp(N=100000, P1=0.05, P2=0.55, P3= 0.5, P4= 0.6, DEFF=3.45, conf=0.99,
cve=0.03, me=0.03, T = 0.5, R = 0.9, plot=TRUE)

#####
# Example with BigLucyT0T1 data #
#####
data(BigLucyT0T1)
attach(BigLucyT0T1)

BigLucyT0 <- BigLucyT0T1[Time == 0,]
BigLucyT1 <- BigLucyT0T1[Time == 1,]
N1 <- table(BigLucyT0$SPAM)[1]
N2 <- table(BigLucyT1$SPAM)[1]
N <- max(N1,N2)
P1 <- prop.table(table(BigLucyT0$ISO))[1]
P2 <- prop.table(table(BigLucyT1$ISO))[1]
P3 <- prop.table(table(BigLucyT0$ISO))[2]
P4 <- prop.table(table(BigLucyT1$ISO))[2]
# The minimum sample size for simple random sampling
ss4ddp(N, P1, P2, P3, P4, conf=0.95, cve=0.05, me=0.03, plot=TRUE)
# The minimum sample size for a complex sampling design
ss4ddp(N, P1, P2, P3, P4, T = 0.5, R = 0.5, conf=0.95, cve=0.05, me=0.03, plot=TRUE)
```

 ss4ddpH

The required sample size for testing a null hypothesis for a double difference of proportions

Description

This function returns the minimum sample size required for testing a null hypothesis regarding a double difference of proportion.

Usage

```
ss4ddpH(
  N,
  P1,
  P2,
  P3,
  P4,
  D,
  DEFF = 1,
  conf = 0.95,
  power = 0.8,
  T = 0,
  R = 1,
  plot = FALSE
)
```

Arguments

N	The maximum population size between the groups (strata) that we want to compare.
P1	The value of the first estimated proportion.
P2	The value of the second estimated proportion.
P3	The value of the third estimated proportion.
P4	The value of the fourth estimated proportion.
D	The minimum effect to test.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
power	The statistical power. By default power = 0.80.
T	The overlap between waves. By default T = 0.
R	The correlation between waves. By default R = 1.
plot	Optionally plot the effect against the sample size.

Details

We assume that it is of interest to test the following set of hypothesis:

$$H_0 : (P_1 - P_2) - (P_3 - P_4) = 0 \quad vs. \quad H_a : (P_1 - P_2) - (P_3 - P_4) = D \neq 0$$

Note that the minimum sample size, restricted to the predefined power β and confidence $1 - \alpha$, is defined by:

$$n = \frac{S^2}{\frac{D^2}{(z_{1-\alpha} + z_\beta)^2} + \frac{S^2}{N}}$$

Where $S^2 = (P_1 * Q_1 + P_2 * Q_2 + P_3 * Q_3 + P_4 * Q_4) * (1 - (T * R)) * DEFF$ and $Q_i = 1 - P_i$ for $i = 1, 2, 3, 4$.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4pH](#)

Examples

```
ss4ddpH(N = 100000, P1 = 0.5, P2 = 0.5, P3 = 0.5, P4 = 0.5, D=0.03)
ss4ddpH(N = 100000, P1 = 0.5, P2 = 0.5, P3 = 0.5, P4 = 0.5, D=0.03, plot=TRUE)
ss4ddpH(N = 100000, P1 = 0.5, P2 = 0.5, P3 = 0.5, P4 = 0.5, D=0.03, DEFF = 2, plot=TRUE)
ss4ddpH(N = 100000, P1 = 0.5, P2 = 0.5, P3 = 0.5, P4 = 0.5,
D=0.03, conf = 0.99, power = 0.9, DEFF = 2, plot=TRUE)
```

```
#####
# Example with BigLucyT0T1 data #
#####
data(BigLucyT0T1)
attach(BigLucyT0T1)
```

```
BigLucyT0 <- BigLucyT0T1[Time == 0,]
BigLucyT1 <- BigLucyT0T1[Time == 1,]
N1 <- table(BigLucyT0$SPAM)[1]
N2 <- table(BigLucyT1$SPAM)[1]
N <- max(N1,N2)
P1 <- prop.table(table(BigLucyT0$ISO))[1]
P2 <- prop.table(table(BigLucyT1$ISO))[1]
P3 <- prop.table(table(BigLucyT0$ISO))[2]
P4 <- prop.table(table(BigLucyT1$ISO))[2]
# The minimum sample size for simple random sampling
ss4ddpH(N, P1, P2, P3, P4, D = 0.05, plot=TRUE)
```



```
# The minimum sample size for a complex sampling design
ss4ddpH(N, P1, P2, P3, P4, D = 0.05, DEFF = 2, T = 0.5, R = 0.5, conf=0.95, plot=TRUE)
```

ss4dm	<i>The required sample size for estimating a single difference of proportions</i>
-------	---

Description

This function returns the minimum sample size required for estimating a single proportion subjecto to predefined errors.

Usage

```
ss4dm(
  N,
  mu1,
  mu2,
  sigma1,
  sigma2,
  DEFF = 1,
  conf = 0.95,
  cve = 0.05,
  rme = 0.03,
  T = 0,
  R = 1,
  plot = FALSE
)
```

Arguments

N	The maximun population size between the groups (strata) that we want to compare.
mu1	The value of the estimated mean of the variable of interes for the first population.
mu2	The value of the estimated mean of the variable of interes for the second population.
sigma1	The value of the estimated variance of the variable of interes for the first population.
sigma2	The value of the estimated mean of a variable of interes for the second population.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95. By default conf = 0.95.
cve	The maximun coeficient of variation that can be allowed for the estimation.
rme	The maximun relative margin of error that can be allowed for the estimation.

T	The overlap between waves. By default T = 0.
R	The correlation between waves. By default R = 1.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

Note that the minimum sample size to achieve a relative margin of error ε is defined by:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Where

$$n_0 = \frac{z_{1-\frac{\alpha}{2}}^2 S^2}{\varepsilon^2 (\mu_1 - \mu_2)^2}$$

and $S^2 = (\sigma_1^2 + \sigma_2^2) * (1 - (T * R)) * DEFF$ Also note that the minimum sample size to achieve a coefficient of variation *cve* is defined by:

$$n = \frac{S^2}{|\bar{y}_1 - \bar{y}_2|^2 cve^2 + \frac{S^2}{N}}$$

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro@gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[e4p](#)

Examples

```
ss4dm(N=100000, mu1=50, mu2=55, sigma1 = 10, sigma2 = 12, cve=0.05, rme=0.03)
ss4dm(N=100000, mu1=50, mu2=55, sigma1 = 10, sigma2 = 12, cve=0.05, rme=0.03, plot=TRUE)
ss4dm(N=100000, mu1=50, mu2=55, sigma1 = 10, sigma2 = 12, DEFF=3.45, conf=0.99, cve=0.03,
      rme=0.03, plot=TRUE)
```

```
#####
# Example with BigLucy data #
#####
data(BigLucy)
attach(BigLucy)

N1 <- table(SPAM)[1]
N2 <- table(SPAM)[2]
N <- max(N1,N2)
```

```

BigLucy.yes <- subset(BigLucy, SPAM == 'yes')
BigLucy.no <- subset(BigLucy, SPAM == 'no')
mu1 <- mean(BigLucy.yes$Income)
mu2 <- mean(BigLucy.no$Income)
sigma1 <- sd(BigLucy.yes$Income)
sigma2 <- sd(BigLucy.no$Income)

# The minimum sample size for simple random sampling
ss4dm(N, mu1, mu2, sigma1, sigma2, DEFF=1, conf=0.99, cve=0.03, rme=0.03, plot=TRUE)
# The minimum sample size for a complex sampling design
ss4dm(N, mu1, mu2, sigma1, sigma2, DEFF=3.45, conf=0.99, cve=0.03, rme=0.03, plot=TRUE)

```

ss4dmH

The required sample size for testing a null hypothesis for a single difference of proportions

Description

This function returns the minimum sample size required for testing a null hypothesis regarding a single difference of proportions.

Usage

```

ss4dmH(
  N,
  mu1,
  mu2,
  sigma1,
  sigma2,
  D,
  DEFF = 1,
  conf = 0.95,
  power = 0.8,
  T = 0,
  R = 1,
  plot = FALSE
)

```

Arguments

N	The maximum population size between the groups (strata) that we want to compare.
mu1	The value of the estimated mean of the variable of interest for the first population.
mu2	The value of the estimated mean of the variable of interest for the second population.
sigma1	The value of the estimated variance of the variable of interest for the first population.

sigma2	The value of the estimated mean of a variable of interest for the second population.
D	The minimum effect to test.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
power	The statistical power. By default power = 0.80.
T	The overlap between waves. By default T = 0.
R	The correlation between waves. By default R = 1.
plot	Optionally plot the effect against the sample size.

Details

We assume that it is of interest to test the following set of hypothesis:

$$H_0 : \mu u_1 - \mu u_2 = 0 \quad vs. \quad H_a : \mu u_1 - \mu u_2 = D \neq 0$$

Note that the minimum sample size, restricted to the predefined power β and confidence $1 - \alpha$, is defined by:

$$n = \frac{S^2}{\frac{D^2}{(z_{1-\alpha} + z_\beta)^2} + \frac{S^2}{N}}$$

where $S^2 = (\sigma_1^2 + \sigma_2^2) * (1 - (T * R)) * DEFF$

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Editorial Universidad Santo Tomás

See Also

[ss4pH](#)

Examples

```
ss4dmH(N = 100000, mu1=50, mu2=55, sigma1 = 10, sigma2 = 12, D=3)
ss4dmH(N = 100000, mu1=50, mu2=55, sigma1 = 10, sigma2 = 12, D=1, plot=TRUE)
ss4dmH(N = 100000, mu1=50, mu2=55, sigma1 = 10, sigma2 = 12, D=0.5, DEFF = 2, plot=TRUE)
ss4dmH(N = 100000, mu1=50, mu2=55, sigma1 = 10, sigma2 = 12, D=0.5, DEFF = 2, conf = 0.99,
power = 0.9, plot=TRUE)
```

```
#####
# Example with BigLucy data #
#####
data(BigLucy)
```

```

attach(BigLucy)

N1 <- table(SPAM)[1]
N2 <- table(SPAM)[2]
N <- max(N1,N2)

BigLucy.yes <- subset(BigLucy, SPAM == 'yes')
BigLucy.no <- subset(BigLucy, SPAM == 'no')
mu1 <- mean(BigLucy.yes$Income)
mu2 <- mean(BigLucy.no$Income)
sigma1 <- sd(BigLucy.yes$Income)
sigma2 <- sd(BigLucy.no$Income)

# The minimum sample size for testing
# H_0: mu_1 - mu_2 = 0 vs. H_a: mu_1 - mu_2 = D = 3
D = 3
ss4dmH(N, mu1, mu2, sigma1, sigma2, D, DEFF = 2, plot=TRUE)

# The minimum sample size for testing
# H_0: mu_1 - mu_2 = 0 vs. H_a: mu_1 - mu_2 = D = 3
D = 3
ss4dmH(N, mu1, mu2, sigma1, sigma2, D, conf = 0.99, power = 0.9, DEFF = 3.45, plot=TRUE)

```

ss4dp

The required sample size for estimating a single difference of proportions

Description

This function returns the minimum sample size required for estimating a single proportion subjecto to predefined errors.

Usage

```

ss4dp(
  N,
  P1,
  P2,
  DEFF = 1,
  conf = 0.95,
  cve = 0.05,
  me = 0.03,
  T = 0,
  R = 1,
  plot = FALSE
)

```

Arguments

N	The maximum population size between the groups (strata) that we want to compare.
P1	The value of the first estimated proportion.
P2	The value of the second estimated proportion.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95. By default conf = 0.95.
cve	The maximum coefficient of variation that can be allowed for the estimation.
me	The maximum margin of error that can be allowed for the estimation.
T	The overlap between waves. By default T = 0.
R	The correlation between waves. By default R = 1.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

Note that the minimum sample size to achieve a particular margin of error ε is defined by:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Where

$$n_0 = \frac{z_{1-\frac{\alpha}{2}}^2 S^2}{\varepsilon^2}$$

and

$$S^2 = (P1 * Q1 + P2 * Q2) * (1 - (T * R)) * DEFF$$

Also note that the minimum sample size to achieve a particular coefficient of variation *cve* is defined by:

$$n = \frac{S^2}{p^2 cve^2 + \frac{S^2}{N}}$$

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Editorial Universidad Santo Tomas

See Also

[e4p](#)

Examples

```

ss4dp(N=100000, P1=0.5, P2=0.55, cve=0.05, me=0.03)
ss4dp(N=100000, P1=0.5, P2=0.55, cve=0.05, me=0.03, plot=TRUE)
ss4dp(N=100000, P1=0.5, P2=0.55, DEFF=3.45, conf=0.99, cve=0.03, me=0.03, plot=TRUE)
ss4dp(N=100000, P1=0.5, P2=0.55, DEFF=3.45, T=0.5, R=0.5, conf=0.99, cve=0.03, me=0.03, plot=TRUE)

#####
# Example with BigLucy data #
#####
data(BigLucy)
attach(BigLucy)

N1 <- table(SPAM)[1]
N2 <- table(SPAM)[2]
N <- max(N1,N2)
P1 <- prop.table(table(SPAM))[1]
P2 <- prop.table(table(SPAM))[2]
# The minimum sample size for simple random sampling
ss4dp(N, P1, P2, DEFF=1, conf=0.99, cve=0.03, me=0.03, plot=TRUE)
# The minimum sample size for a complex sampling design
ss4dp(N, P1, P2, DEFF=3.45, conf=0.99, cve=0.03, me=0.03, plot=TRUE)

```

ss4dpH

The required sample size for testing a null hypothesis for a single difference of proportions

Description

This function returns the minimum sample size required for testing a null hypothesis regarding a single proportion.

Usage

```

ss4dpH(
  N,
  P1,
  P2,
  D,
  DEFF = 1,
  conf = 0.95,
  power = 0.8,
  T = 0,
  R = 1,
  plot = FALSE
)

```

Arguments

N	The maximum population size between the groups (strata) that we want to compare.
P1	The value of the first estimated proportion.
P2	The value of the second estimated proportion.
D	The minimum effect to test.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
power	The statistical power. By default power = 0.80.
T	The overlap between waves. By default T = 0.
R	The correlation between waves. By default R = 1.
plot	Optionally plot the effect against the sample size.

Details

We assume that it is of interest to test the following set of hypothesis:

$$H_0 : P_1 - P_2 = 0 \quad vs. \quad H_a : P_1 - P_2 = D \neq 0$$

Note that the minimum sample size, restricted to the predefined power β and confidence $1 - \alpha$, is defined by:

$$n = \frac{S^2}{\frac{D^2}{(z_{1-\alpha} + z_\beta)^2} + \frac{S^2}{N}}$$

Where $S^2 = (P1 * Q1 + P2 * Q2) * (1 - (T * R)) * DEFF$ and $Q_i = 1 - P_i$ for $i = 1, 2$.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Editorial Universidad Santo Tomas

See Also

[ss4pH](#)

Examples

```
ss4dpH(N = 100000, P1 = 0.5, P2 = 0.55, D=0.03)
ss4dpH(N = 100000, P1 = 0.5, P2 = 0.55, D=0.03, plot=TRUE)
ss4dpH(N = 100000, P1 = 0.5, P2 = 0.55, D=0.03, DEFF = 2, plot=TRUE)
ss4dpH(N = 100000, P1 = 0.5, P2 = 0.55, D=0.03, conf = 0.99, power = 0.9, DEFF = 2, plot=TRUE)
```



```
#####
# Example with BigLucy data #
#####
data(BigLucy)
attach(BigLucy)

N1 <- table(SPAM)[1]
N2 <- table(SPAM)[2]
N <- max(N1,N2)
P1 <- prop.table(table(SPAM))[1]
P2 <- prop.table(table(SPAM))[2]

# The minimum sample size for testing
# H_0: P_1 - P_2 = 0 vs. H_a: P_1 - P_2 = D = 0.05
D = 0.05
ss4dpH(N, P1, P2, D, DEFF = 2, plot=TRUE)

# The minimum sample size for testing
# H_0: P - P_0 = 0 vs. H_a: P - P_0 = D = 0.02
D = 0.01
ss4dpH(N, P1, P2, D, conf = 0.99, power = 0.9, DEFF = 3.45, plot=TRUE)
```

ss4HHSm

Sample Sizes for Household Surveys in Two-Stages for Estimating Single Means

Description

This function computes a grid of possible sample sizes for estimating single means under two-stage sampling designs.

Usage

```
ss4HHSm(N, M, rho, mu, sigma, delta, conf, m)
```

Arguments

N	The population size.
M	Number of clusters in the population.
rho	The Intraclass Correlation Coefficient.
mu	The value of the estimated mean of a variable of interest.
sigma	The value of the estimated standard deviation of a variable of interest.
delta	The maximum margin of error that can be allowed for the estimation.
conf	The statistical confidence. By default conf = 0.95.
m	(vector) Number of households selected within PSU.

Details

In two-stage (2S) sampling, the design effect is defined by

$$DEFF = 1 + (\bar{m} - 1)\rho$$

Where ρ is defined as the intraclass correlation coefficient, \bar{m} is the average sample size of units selected inside each cluster. The relationship of the full sample size of the two stage design (2S) with the simple random sample (SI) design is given by

$$n_{2S} = n_{SI} * DEFF$$

Value

This function returns a grid of possible sample sizes. The first column represent the design effect, the second column is the number of clusters to be selected, the third column is the number of units to be selected inside the clusters, and finally, the last column indicates the full sample size induced by this particular strategy.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ICC](#)

Examples

```
ss4HHSm(N = 50000000, M = 3000, rho = 0.034,
        mu = 10, sigma = 2, delta = 0.03, conf = 0.95,
        m = c(5:15))
```

```
#####
# Example with BigCity data      #
# Sample size for the estimation #
# of the unemployment rate      #
#####
```

```
library(TeachingSampling)
data(BigCity)
```

```
BigCity1 <- BigCity %>%
  group_by(HHID) %>%
  summarise(IncomeHH = sum(Income),
            PSU = unique(PSU))
```

```

summary(BigCity1$IncomeHH)
mean(BigCity1$IncomeHH)
sd(BigCity1$IncomeHH)

N <- nrow(BigCity)
M <- length(unique(BigCity$PSU))
rho <- ICC(BigCity1$IncomeHH, BigCity1$PSU)$ICC
mu <- mean(BigCity1$IncomeHH)
sigma <- sd(BigCity1$IncomeHH)
delta <- 0.05
conf <- 0.95
m <- c(5:15)
ss4HHSm(N, M, rho, mu, sigma, delta, conf, m)

```

ss4HHSp

Sample Sizes for Household Surveys in Two-Stages for Estimating Single Proportions

Description

This function computes a grid of possible sample sizes for estimating single proportions under two-stage sampling designs.

Usage

```
ss4HHSp(N, M, r, b, rho, P, delta, conf, m)
```

Arguments

N	The population size.
M	Number of clusters in the population.
r	Percentage of people within the subpopulation of interest.
b	Average household size (number of members).
rho	The Intraclass Correlation Coefficient.
P	The value of the estimated proportion.
delta	The maximum margin of error that can be allowed for the estimation.
conf	The statistical confidence. By default conf = 0.95.
m	(vector) Number of households selected within PSU.

Details

In two-stage (2S) sampling, the design effect is defined by

$$DEFF = 1 + (\bar{m} - 1)\rho$$

Where ρ is defined as the intraclass correlation coefficient, \bar{m} is the average sample size of units selected inside each cluster. The relationship of the full sample size of the two stage design (2S) with the simple random sample (SI) design is given by

$$n_{2S} = n_{SI} * DEFF$$

Value

This function returns a grid of possible sample sizes. The first column represent the design effect, the second column is the number of clusters to be selected, the third column is the number of units to be selected inside the clusters, and finally, the last column indicates the full sample size induced by this particular strategy.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ICC](#)

Examples

```
ss4HHSp(N = 50000000, M = 3000, r = 1, b = 3.5,
rho = 0.034, P = 0.05, delta = 0.05, conf = 0.95,
m = c(5:15))
```

```
#####
# Example with BigCity data      #
# Sample size for the estimation #
# of the unemployment rate      #
#####
```

```
library(TeachingSampling)
data(BigCity)
```

```
BigCity1 <- BigCity[!is.na(BigCity$Employment), ]
summary(BigCity1$Employment)
BigCity1$Unemp <- Domains(BigCity1$Employment)[, 1]
BigCity1$Active <- Domains(BigCity1$Employment)[, 1] +
```

```

Domains(BigCity1$Employment)[, 3]

N <- nrow(BigCity)
M <- length(unique(BigCity$PSU))
r <- sum(BigCity1$Active)/N
b <- N/length(unique(BigCity$HHID))
rho <- ICC(BigCity1$Unemp, BigCity1$PSU)$ICC
P <- sum(BigCity1$Unemp)/sum(BigCity1$Active)
delta <- 0.05
conf <- 0.95
m <- c(5:15)
ss4HHSp(N, M, r, b, rho, P, delta, conf, m)

```

ss4m

The required sample size for estimating a single mean

Description

This function returns the minimum sample size required for estimating a single mean subject to predefined errors.

Usage

```

ss4m(
  N,
  mu,
  sigma,
  DEFF = 1,
  conf = 0.95,
  error = "cve",
  delta = 0.03,
  plot = FALSE
)

```

Arguments

N	The population size.
mu	The value of the estimated mean of a variable of interest.
sigma	The value of the estimated standard deviation of a variable of interest.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95. By default conf = 0.95.
error	The type of error you want to minimize.
delta	The magnitude of the error you want to minimize.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

Note that the minimum sample size to achieve a relative margin of error ε is defined by:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Where

$$n_0 = \frac{z_{1-\frac{\alpha}{2}}^2 S^2}{\varepsilon^2 \mu^2}$$

and

$$S^2 = \sigma^2 DEFF$$

Also note that the minimum sample size to achieve a coefficient of variation *cve* is defined by:

$$n = \frac{S^2}{\bar{y}_U^2 cve^2 + \frac{S^2}{N}}$$

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[e4p](#)

Examples

```
ss4m(N=10000, mu=10, sigma=2, DEFF = 2, error = "cve", delta = 0.03, plot=TRUE)
ss4m(N=10000, mu=10, sigma=2, DEFF = 2, error = "me", delta = 1, plot=TRUE)
ss4m(N=10000, mu=10, sigma=2, DEFF = 2, error = "rme", delta = 0.03, plot=TRUE)
```

```
#####
# Example with Lucy data #
#####
```

```
data(Lucy)
attach(Lucy)
N <- nrow(Lucy)
mu <- mean(Income)
sigma <- sd(Income)
# The minimum sample size for simple random sampling
ss4m(N, mu, sigma, DEFF=1, conf=0.95, error = "rme", delta = 0.03, plot=TRUE)
# The minimum sample size for a complex sampling design
ss4m(N, mu, sigma, DEFF=1, conf=0.95, error = "me", delta = 5, plot=TRUE)
# The minimum sample size for a complex sampling design
ss4m(N, mu, sigma, DEFF=3.45, conf=0.95, error = "rme", delta = 0.03, plot=TRUE)
```

 ss4mH

The required sample size for testing a null hypothesis for a single mean

Description

This function returns the minimum sample size required for testing a null hypothesis regarding a single mean

Usage

```
ss4mH(N, mu, mu0, sigma, DEFF = 1, conf = 0.95, power = 0.8, plot = FALSE)
```

Arguments

N	The population size.
mu	The population mean of the variable of interest.
mu0	The value to test for the single mean.
sigma	The population variance of the variable of interest.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
power	The statistical power. By default power = 0.80.
plot	Optionally plot the effect against the sample size.

Details

We assume that it is of interest to test the following set of hypothesis:

$$H_0 : \mu - \mu_0 = 0 \quad \text{vs.} \quad H_a : \mu - \mu_0 = D \neq 0$$

Note that the minimum sample size, restricted to the predefined power β and confidence $1 - \alpha$, is defined by:

$$n = \frac{S^2}{\frac{D^2}{(z_{1-\alpha} + z_\beta)^2} + \frac{S^2}{N}}$$

Where $S^2 = \sigma^2 * DEFF$ and σ^2 is the population variance of the variable of interest.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Editorial Universidad Santo Tomas

DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95. By default conf = 0.95.
error	The type of error you want to minimize.
delta	The magnitude of the error you want to minimize.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

Note that the minimum sample size to achieve a particular margin of error ε is defined by:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Where

$$n_0 = \frac{z_{1-\frac{\alpha}{2}}^2 S^2}{\varepsilon^2}$$

and

$$S^2 = P(1 - P)DEFF$$

Also note that the minimum sample size to achieve a particular coefficient of variation *cve* is defined by:

$$n = \frac{S^2}{P^2 cve^2 + \frac{S^2}{N}}$$

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[e4p](#)

Examples

```
ss4p(N=10000, P=0.05, error = "cve", delta=0.05, DEFF = 1, conf = 0.95, plot=TRUE)
ss4p(N=10000, P=0.05, error = "me", delta=0.05, DEFF = 1, conf = 0.95, plot=TRUE)
ss4p(N=10000, P=0.5, error = "rme", delta=0.05, DEFF = 1, conf = 0.95, plot=TRUE)
```

```
#####
# Example with Lucy data #
#####
```

```
data(Lucy)
attach(Lucy)
```

```

N <- nrow(Lucy)
P <- prop.table(table(SPAM))[1]
# The minimum sample size for simple random sampling
ss4p(N, P, DEFF=3.45, conf=0.95, error = "cve", delta = 0.03, plot=TRUE)
# The minimum sample size for a complex sampling design
ss4p(N, P, DEFF=3.45, conf=0.95, error = "rme", delta = 0.03, plot=TRUE)
# The minimum sample size for a complex sampling design
ss4p(N, P, DEFF=3.45, conf=0.95, error = "me", delta = 0.03, plot=TRUE)

```

ss4pH	<i>The required sample size for testing a null hypothesis for a single proportion</i>
-------	---

Description

This function returns the minimum sample size required for testing a null hypothesis regarding a single proportion.

Usage

```
ss4pH(N, p, p0, DEFF = 1, conf = 0.95, power = 0.8, plot = FALSE)
```

Arguments

N	The population size.
p	The value of the estimated proportion.
p0	The value to test for the single proportion.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
power	The statistical power. By default power = 0.80.
plot	Optionally plot the effect against the sample size.

Details

We assume that it is of interest to test the following set of hypothesis:

$$H_0 : P - P_0 = 0 \quad \text{vs.} \quad H_a : P - P_0 = D \neq 0$$

Note that the minimum sample size, restricted to the predefined power β and confidence $1 - \alpha$, is defined by:

$$n = \frac{S^2}{\frac{D^2}{(z_{1-\alpha} + z_\beta)^2} + \frac{S^2}{N}}$$

Where

$$S^2 = p(1 - p)DEFF$$

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[e4p](#)

Examples

```

ss4pH(N = 10000, p = 0.5, p0 = 0.55)
ss4pH(N = 10000, p = 0.5, p0 = 0.55, plot=TRUE)
ss4pH(N = 10000, p = 0.5, p0 = 0.55, DEFF = 2, plot=TRUE)
ss4pH(N = 10000, p = 0.5, p0 = 0.55, conf = 0.99, power = 0.9, DEFF = 2, plot=TRUE)

#####
# Example with BigLucy data #
#####
data(BigLucy)
attach(BigLucy)

N <- nrow(BigLucy)
p <- prop.table(table(SPAM))[1]

# The minimum sample size for testing
# H_0: P - P_0 = 0 vs. H_a: P - P_0 = D = 0.1
D = 0.1
p0 = p - D
ss4pH(N, p, p0, conf = 0.99, power = 0.9, DEFF = 2, plot=TRUE)

# The minimum sample size for testing
# H_0: P - P_0 = 0 vs. H_a: P - P_0 = D = 0.02
D = 0.02
p0 = p - D
ss4pH(N, p, p0, conf = 0.99, power = 0.9, DEFF = 3.45, plot=TRUE)

```

ss4pLN

The required sample size for estimating a single proportion based on a logarithmic transformation of the estimated proportion

Description

This function returns the minimum sample size required for estimating a single proportion subjecto to predefined errors.

Usage

```
ss4pLN(N, P, DEFF = 1, cve = 0.05, plot = FALSE)
```

Arguments

N	The population size.
P	The value of the estimated proportion.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
cve	The maximum coefficient of variation that can be allowed for the estimation.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

As for low proportions, the coefficient of variation tends to infinity, it is customary to use a symmetrical transformation of this measure (based on the relative standard error RSE) to report the uncertainty of the estimation. This way, if $p \leq 0.5$, the transformed CV will be:

$$RSE(-\ln(p)) = \frac{SE(p)}{-\ln(p) * p}$$

Otherwise, when $p > 0.5$, the transformed CV will be:

$$RSE(-\ln(1-p)) = \frac{SE(p)}{-\ln(1-p) * (1-p)}$$

Note that, when $p \leq 0.5$ the minimum sample size to achieve a particular coefficient of variation *cve* is defined by:

$$n = \frac{S^2}{P^2 cve^2 + \frac{S^2}{N}}$$

When $p > 0.5$ the minimum sample size to achieve a particular coefficient of variation *cve* is defined by:

$$n = \frac{S^2}{P^2 cve^2 + \frac{S^2}{N}}$$

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Editorial Universidad Santo Tomas

See Also

[ss4p](#)

Examples

```

ss4pLN(N=10000, P=0.8, cve=0.10)
ss4pLN(N=10000, P=0.2, cve=0.10)
ss4pLN(N=10000, P=0.7, cve=0.05, plot=TRUE)
ss4pLN(N=10000, P=0.3, cve=0.05, plot=TRUE)
ss4pLN(N=10000, P=0.05, DEFF=3.45, cve=0.03, plot=TRUE)
ss4pLN(N=10000, P=0.95, DEFF=3.45, cve=0.03, plot=TRUE)

```

```

#####
# Example with Lucy data #
#####

```

```

data(Lucy)
attach(Lucy)
N <- nrow(Lucy)
P <- prop.table(table(SPAM))[1]
# The minimum sample size for simple random sampling
ss4pLN(N, P, DEFF=1, cve=0.03, plot=TRUE)
# The minimum sample size for a complex sampling design
ss4pLN(N, P, DEFF=3.45, cve=0.03, plot=TRUE)

```

ss4S2

The required sample size for estimating a single variance

Description

This function returns the minimum sample size required for estimating a single variance subjecto to predefined errors.

Usage

```
ss4S2(N, K = 0, DEFF = 1, conf = 0.95, cve = 0.05, me = 0.03, plot = FALSE)
```

Arguments

N	The population size.
K	The population excess kurtosis of the variable in the population.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95. By default conf = 0.95.
cve	The maximun coeficient of variation that can be allowed for the estimation.
me	The maximun margin of error that can be allowed for the estimation.
plot	Optionally plot the errors (cve and margin of error) against the sample size.

Details

Note that the minimum sample size to achieve a particular relative margin of error ε is defined by:

$$n = \frac{n_0}{\frac{(N-1)^3}{N^2(N*K+2N+2)} + \frac{n_0}{N}}$$

Where

$$n_0 = \frac{z_{1-\frac{\alpha}{2}}^2 * DEFF}{\varepsilon^2}$$

Also note that the minimum sample size to achieve a particular coefficient of variation cve is defined by:

$$n = \frac{N^2(N * K + 2N + 2) * DEFF}{cve^2 * (N - 1)^3 + N(N * K + 2N + 2) * DEFF}$$

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[e4p](#)

Examples

```
ss4S2(N = 10000, K = 0, cve = 0.05, me = 0.03)
ss4S2(N = 10000, K = 1, cve = 0.05, me = 0.03)
ss4S2(N = 10000, K = 1, cve = 0.05, me = 0.05, DEFF = 2)
ss4S2(N = 10000, K = 1, cve = 0.05, me = 0.03, plot = TRUE)

#####
# Example with BigLucy data #
#####

data(BigLucy)
attach(BigLucy)
N <- nrow(BigLucy)
K <- kurtosis(BigLucy$Income)
# The minimum sample size for simple random sampling
ss4S2(N, K, DEFF=1, conf=0.99, cve=0.03, me=0.1, plot=TRUE)
# The minimum sample size for a complex sampling design
ss4S2(N, K, DEFF=3.45, conf=0.99, cve=0.03, me=0.1, plot=TRUE)
```

ss4S2H	<i>The required sample size for testing a null hypothesis for a single variance</i>
--------	---

Description

This function returns the minimum sample size required for testing a null hypothesis regarding a single variance

Usage

```
ss4S2H(N, S2, S20, K = 0, DEFF = 1, conf = 0.95, power = 0.8, plot = FALSE)
```

Arguments

N	The population size.
S2	The value of the estimated variance
S20	The value to test for the single variance
K	The excess kurtosis of the variable in the population.
DEFF	The design effect of the sample design. By default DEFF = 1, which corresponds to a simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
power	The statistical power. By default power = 0.80.
plot	Optionally plot the effect against the sample size.

Details

We assume that it is of interest to test the following set of hypothesis:

$$H_0 : P - P_0 = 0 \quad vs. \quad H_a : P - P_0 = D > 0$$

Note that the minimum sample size, restricted to the predefined power β and confidence $1 - \alpha$, is defined by:

$$n = \frac{S2^2}{\frac{D^2}{(z_{1-\alpha} + z_\beta)^2} \frac{(N-1)^3}{N^2(N*K+2N+2)} + \frac{S2^2}{N}}$$

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseno de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also[e4p](#)**Examples**

```

ss4S2H(N = 10000, S2 = 120, S20 = 110, K = 0)
ss4S2H(N = 10000, S2 = 120, S20 = 110, K = 2, DEFF = 2, power = 0.9)
ss4S2H(N = 10000, S2 = 120, S20 = 110, K = 2, DEFF = 2, power = 0.8, plot = TRUE)

#####
# Example with BigLucy data #
#####
data(BigLucy)
attach(BigLucy)
N <- nrow(BigLucy)
S2 <- var(BigLucy$Income)

# The minimum sample size for testing
# H_0: S2 - S2_0 = 0 vs. H_a: S2 - S2_0 = D = 8000
D = 8000
S20 = S2 - D
K <- kurtosis(BigLucy$Income)
ss4S2H(N, S2, S20, K, DEFF=1, conf = 0.99, power = 0.8, plot=TRUE)

```

ss4stm

*Sample Size for Estimation of Means in Stratified Sampling***Description**

This function computes the minimum sample size required for estimating a single mean, in a stratified sampling, subject to predefined errors.

Usage

```
ss4stm(Nh, muh, sigmah, DEFFh = 1, conf = 0.95, rme = 0.03)
```

Arguments

Nh	Vector. The population size for each stratum.
muh	Vector. The means of the variable of interest in each stratum.
sigmah	Vector. The standard deviation of the variable of interest in each stratum.
DEFFh	Vector. The design effect of the sample design in each stratum. By default DEFFh = 1, which corresponds to a stratified simple random sampling design.
conf	The statistical confidence. By default conf = 0.95.
rme	The maximum relative margin of error that can be allowed for the estimation.

Details

Let assume that the population U is partitioned in H strata. Under a stratified sampling, the necessary sample size to achieve a relative margin of error ε is defined by:

$$n = \frac{(\sum_{h=1}^H w_h S_h)^2}{\frac{\varepsilon^2}{z_{1-\frac{\alpha}{2}}^2} + \frac{\sum_{h=1}^H w_h S_h^2}{N}}$$

Where

$$S_h^2 = DEFF_h \sigma_h^2$$

Then, the required sample size in each stratum is given by:

$$n_h = n \frac{w_h S_h}{\sum_{h=1}^H w_h S_h}$$

Value

The required sample size for the sample and the required sample size per stratum.

Author(s)

Hugo Andres Gutierrez Rojas <hagutierrezro at gmail.com>

References

Gutierrez, H. A. (2009), *Estrategias de muestreo: Diseño de encuestas y estimacion de parametros*. Editorial Universidad Santo Tomas

See Also

[ss4m](#)

Examples

```
Nh <- c(15000, 10000, 5000)
muh <- c(300, 200, 100)
sigmah <- c(200, 100, 20)
DEFFh <- c(1, 1.2, 1.5)

ss4stm(Nh, muh, sigmah, rme=0.03)
ss4stm(Nh, muh, sigmah, conf = 0.99, rme=0.03)
ss4stm(Nh, muh, sigmah, DEFFh, conf= 0.99, rme=0.03)

#####
# Example with Lucy data #
#####
data(Lucy)
attach(Lucy)
```

```
Strata <- as.factor(paste(Zone, Level))
levels(Strata)

Nh <- summary(Strata)
muh <- tapply(Income, Strata, mean)
sigmah <- tapply(Income, Strata, sd)

ss4stm(Nh, muh, sigmah, DEFFh=1, conf = 0.95, rme = 0.03)
ss4stm(Nh, muh, sigmah, DEFFh=1.5, conf = 0.95, rme = 0.03)

#####
# Example with BigLucy data #
#####
data(BigLucy)
attach(BigLucy)

Nh <- summary(Zone)
muh <- tapply(Income, Zone, mean)
sigmah <- tapply(Income, Zone, sd)

ss4stm(Nh, muh, sigmah, DEFFh=1, conf = 0.95, rme = 0.03)
ss4stm(Nh, muh, sigmah, DEFFh=1.5, conf = 0.95, rme = 0.03)
```

Index

*Topic **datasets**

BigLucyT0T1, [12](#)

b4ddm, [2](#)

b4ddp, [4](#)

b4dm, [5](#)

b4dp, [7](#)

b4m, [8](#)

b4p, [9](#)

b4S2, [10](#)

BigLucyT0T1, [12](#)

DEFF, [13](#)

e4ddm, [15](#)

e4ddp, [17](#)

e4dm, [18](#)

e4dp, [19](#)

e4m, [21](#)

e4p, [22](#), [33](#), [42](#), [46](#), [54](#), [56](#), [57](#), [59](#), [62](#), [64](#)

e4S2, [23](#)

ICC, [24](#), [28](#), [31](#), [50](#), [52](#)

ss2s4m, [26](#)

ss2s4p, [30](#)

ss4ddm, [31](#)

ss4ddmH, [34](#)

ss4ddp, [36](#)

ss4ddpH, [39](#)

ss4dm, [41](#)

ss4dmH, [43](#)

ss4dp, [38](#), [45](#)

ss4dpH, [47](#)

ss4HHSm, [49](#)

ss4HHSp, [51](#)

ss4m, [53](#), [65](#)

ss4mH, [55](#)

ss4p, [4–6](#), [8–11](#), [16](#), [18–20](#), [22–25](#), [56](#), [60](#)

ss4pH, [35](#), [40](#), [44](#), [48](#), [58](#)

ss4pLN, [59](#)

ss4S2, [61](#)

ss4S2H, [63](#)

ss4stm, [64](#)