Package 'skewsamp'

December 16, 2021

Title Estimate Sample Sizes for Group Comparisons with Skewed Distributions

Version 1.0.0

Description Estimate necessary sample sizes for comparing the location of data from two groups or categories when the distribution of the data is skewed. The package offers a non-parametric method for a Wilcoxon Mann-Whitney test of location shift as well as methods for several generalized linear models, for instance, Gamma regression.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.1.1

Imports stats

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

NeedsCompilation no

Author Johannes Brachem [cre, aut], Dominik Strache [aut]

Maintainer Johannes Brachem <jbrachem@posteo.de>

Repository CRAN

Date/Publication 2021-12-16 21:00:08 UTC

R topics documented:

---

demp *Empirical probability density function (EPDF)*

Description

Empirical probability density function based on a sample of observations, as described by Chakraborti (2006).

Usage

```
demp(x, sample)
```

Arguments

x	numeric vector of values to evaluate
sample	numeric vector of sample values to base the EPDF on

Value

numeric vector of density values based on the EPDF

References

Chakraborti, S., Hong, B., & Van De Wiel, M. A. (2006). A note on sample size determination for a nonparametric test of location. Technometrics, 48(1), 88–94. https://doi.org/10.1198/004017005000000193

Examples

```
x <- 1:5
demp(1, x)
```

n_binom *Calculate sample size for binomial distribution*

Description

Estimation of required sample size as given by Cundill & Alexander (2015).

Usage

```
n_binom(
  p0,
  effect,
  size = 1,
  alpha = 0.05,
  power = 0.9,
  q = 0.5,
  link = c("logit", "identity"),
  two_sided = TRUE
)
```

Arguments

p0	probability of success in group0
effect	Effect size, $1 - (\mu_1/\mu_0)$, where μ_0 is the mean in the control group (mean0) and μ_1 is the mean in the treatment group.
size	number of trials (greater than zero)
alpha	Type I error rate
power	1 - Type II error rate
q	Proportion of observations allocated to the control group
link	Link function to use. Currently implement: 'log' and 'identity'
two_sided	logical, if TRUE the sample size will be calculated for a two-sided test. Otherwise, the sample size will be calculated for a one-sided test.

Value

Returns an object of class "sample_size". It contains the following components:

N	the total sample size
n0	sample size in Group 0 (control group)
n1	sample size in Group 1 (treatment group)
two_sided	logical, TRUE, if the estimated sample size refers to a two-sided test
alpha	type I error rate used in sample size estimation
power	target power used in sample size estimation

effect	effect size used in sample size estimation
effect_type	short description of the type of effect size
comment	additional comment, if there is any
call	the matched call.

References

Cundill, B., & Alexander, N. D. E. (2015). Sample size calculations for skewed distributions. *BMC Medical Research Methodology*, 15(1), 1–9. https://doi.org/10.1186/s12874-015-0023-0

Examples

```
n_binom(p0 = 0.5, effect = 0.3)
```

---

n_gamma *Calculate sample size for gamma distribution*

Description

Estimation of required sample size as given by Cundill & Alexander (2015).

Usage

```
n_gamma(
  mean0,
  effect,
  shape0,
  shape1 = shape0,
  alpha = 0.05,
  power = 0.9,
  q = 0.5,
  link = c("log", "identity"),
  two_sided = TRUE
)
```

Arguments

mean0	Mean in control group
effect	Effect size, $1 - (\mu_1/\mu_0)$, where μ_0 is the mean in the control group (mean0) and μ_1 is the mean in the treatment group.
shape0	Shape parameter in control group
shape1	Shape parameter in treatment group. Defaults to shape0, because GLM assumes equal shape across groups.
alpha	Type I error rate
power	1 - Type II error rate

q	Proportion of observations allocated to the control group
link	Link function to use. Currently implement: 'log' and 'identity'
two_sided	logical, if TRUE the sample size will be calculated for a two-sided test. Otherwise, the sample size will be calculated for a one-sided test.

Value

Returns an object of class "sample_size". It contains the following components:

N	the total sample size
n0	sample size in Group 0 (control group)
n1	sample size in Group 1 (treatment group)
two_sided	logical, TRUE, if the estimated sample size refers to a two-sided test
alpha	type I error rate used in sample size estimation
power	target power used in sample size estimation
effect	effect size used in sample size estimation
effect_type	short description of the type of effect size
comment	additional comment, if there is any
call	the matched call.

References

Cundill, B., & Alexander, N. D. E. (2015). Sample size calculations for skewed distributions. *BMC Medical Research Methodology*, 15(1), 1–9. https://doi.org/10.1186/s12874-015-0023-0

Examples

```
n_gamma(mean0 = 8.46, effect = 0.7, shape0 = 0.639,
           alpha = 0.05, power = 0.9)
```

---

n_glm *Calculate sample size for a group comparison via generalized linear models*

Description

Estimation of required sample size as given by Cundill & Alexander (2015).

Usage

```
n_glm(
  mean0,
  mean1,
  dispersion0,
  dispersion1,
  alpha,
  power,
  link_fun = function(mu) NULL,
  variance_fun = function(mu, dispersion) NULL,
  dmu_deta_fun = function(mu) NULL,
  q
)
```

Arguments

`mean0`	Mean in control group
`mean1`	Mean in treatment group
`dispersion0`	Dispersion parameter in control group
`dispersion1`	Dispersion parameter in treatment group.
`alpha`	Type I error rate
`power`	1 - Type II error rate
`link_fun`	function object, the link function to create the response η.
`variance_fun`	function object, function for computing the variance based on a mean and a dispersion parameter
`dmu_deta_fun`	function object, derivative of the original mean with respect to the link: $d\mu/d\eta$.
`q`	Number between 0 and 1, the proportion of observations allocated to the control group

Value

Total sample size (numeric)

References

Cundill, B., & Alexander, N. D. E. (2015). Sample size calculations for skewed distributions. *BMC Medical Research Methodology*, 15(1), 1–9. https://doi.org/10.1186/s12874-015-0023-0

`n_locshift` *Estimate N on the basis of two pilot samples.*

Description

Estimation as described by Chakraborti, Hong, & van de Wiel (2006).

Usage

```
n_locshift(s1, s2, delta, alpha = 0.05, power = 0.9, q = 0.5)
```

Arguments

`s1, s2`	pilot samples
`delta`	numeric value, location shift parameter δ
`alpha`	type-I error probability
`power`	1 - type-II error probability, the desired statistical power
`q`	size of group0 relative to total sample size.

Details

WARNING: Note that the estimation has high variability due to its dependence on pilot samples. The smaller the pilot sample, the more uncertain is the estimation of the required sample size. In a simulation study, we found that the method may also be inaccurate on average, depending on the investigated data.

Value

Returns an object of class `"sample_size"`. It contains the following components:

`N`	the total sample size
`n0`	sample size in Group 0 (control group)
`n1`	sample size in Group 1 (treatment group)
`two_sided`	logical, `TRUE`, if the estimated sample size refers to a two-sided test
`alpha`	type I error rate used in sample size estimation
`power`	target power used in sample size estimation
`effect`	effect size used in sample size estimation
`effect_type`	short description of the type of effect size
`comment`	additional comment, if there is any
`call`	the matched call.

References

Chakraborti, S., Hong, B., & van de Wiel, M. A. (2006). A note on sample size determination for a nonparametric test of location. Technometrics, 48(1), 88–94. https://doi.org/10.1198/004017005000000193

Examples

```
n_locshift(s1 = rexp(10), s2 = rexp(10),
           alpha = 0.05, power = 0.9, delta = 0.35)
```

`n_locshift_bound` *Compute an upper bound the sample size based on two pilot samples.*

Description

Based on the procedure described by Chakraborti, Hong, & van de Wiel (2006)

Usage

```
n_locshift_bound(
  s1,
  s2,
  delta,
  alpha = 0.05,
  power = 0.9,
  quantile = 0.9,
  n_resamples = 500,
  q = 0.5
)
```

Arguments

`s1, s2`	Pilot samples
`delta`	numeric value, location shift parameter δ
`alpha`	Type I error probability
`power`	1 - Type II error probability, the desired statistical power
`quantile`	Quantile to use as the upper bound.
`n_resamples`	number of resamples to use in bootstrapping
`q`	size of group0 relative to total sample size.

Details

WARNING: Note that the underlying estimation has high variability due to its dependence on pilot samples. The smaller the pilot sample, the more uncertain is the estimation of the required sample size. In a simulation study, we found that the underlying method may also be inaccurate on average, depending on the investigated data.

Value

Returns an object of class "sample_size". It contains the following components:

n	the total sample size
n0	sample size in Group 0 (control group)
n1	sample size in Group 1 (treatment group)
two_sided	logical, TRUE, if the estimated sample size refers to a two-sided test
alpha	type I error rate used in sample size estimation
power	target power used in sample size estimation
effect	effect size used in sample size estimation
effect_type	short description of the type of effect size
comment	additional comment, if there is any
call	the matched call.

References

Chakraborti, S., Hong, B., & van de Wiel, M. A. (2006). A note on sample size determination for a nonparametric test of location. Technometrics, 48(1), 88–94. https://doi.org/10.1198/004017005000000193

Examples

```
## Not run:
n_locshift_bound(s1 = rexp(10), s2 = rexp(10),
           delta = 0.35, alpha = 0.05, power = 0.9, n_resamples = 5)
## End(Not run)
```

---

n_negbinom *Calculate sample size for negative binomial distribution*

Description

Estimation of required sample size as given by Cundill & Alexander (2015).

Usage

```
n_negbinom(
  mean0,
  effect,
  dispersion0,
  dispersion1 = dispersion0,
  alpha = 0.05,
  power = 0.9,
  q = 0.5,
  link = c("log", "identity"),
  two_sided = TRUE
)
```

Arguments

mean0	Mean in control group
effect	Effect size, $1 - (\mu_1/\mu_0)$, where μ_0 is the mean in the control group (mean0) and μ_1 is the mean in the treatment group.
dispersion0	Dispersion parameter in control group
dispersion1	Dispersion parameter in treatment group. Defaults to shape0, because GLM assumes equal shape across groups.
alpha	Type I error rate
power	1 - Type II error rate
q	Proportion of observations allocated to the control group
link	Link function to use. Currently implement: 'log' and 'identity'
two_sided	logical, if TRUE the sample size will be calculated for a two-sided test. Otherwise, the sample size will be calculated for a one-sided test.

Value

Returns an object of class "sample_size". It contains the following components:

N	the total sample size
n0	sample size in Group 0 (control group)
n1	sample size in Group 1 (treatment group)
two_sided	logical, TRUE, if the estimated sample size refers to a two-sided test
alpha	type I error rate used in sample size estimation
power	target power used in sample size estimation
effect	effect size used in sample size estimation
effect_type	short description of the type of effect size
comment	additional comment, if there is any
call	the matched call.

References

Cundill, B., & Alexander, N. D. E. (2015). Sample size calculations for skewed distributions. *BMC Medical Research Methodology*, 15(1), 1–9. https://doi.org/10.1186/s12874-015-0023-0

Examples

```
n_negbinom(mean0 = 71.4, effect = 0.7, dispersion0 = 0.33,
           alpha = 0.05, power = 0.9)
```

`n_poisson` *Calculate sample size for poisson distribution*

Description

Estimation of required sample size as given by Cundill & Alexander (2015).

Usage

```
n_poisson(
  mean0,
  effect,
  alpha = 0.05,
  power = 0.9,
  q = 0.5,
  link = c("log", "identity"),
  two_sided = TRUE
)
```

Arguments

`mean0`	Mean in control group
`effect`	Effect size, $1 - (\mu_1/\mu_0)$, where μ_0 is the mean in the control group (`mean0`) and μ_1 is the mean in the treatment group.
`alpha`	Type I error rate
`power`	1 - Type II error rate
`q`	Proportion of observations allocated to the control group
`link`	Link function to use. Currently implement: 'log' and 'identity'
`two_sided`	logical, if `TRUE` the sample size will be calculated for a two-sided test. Otherwise, the sample size will be calculated for a one-sided test.

Value

Returns an object of class `"sample_size"`. It contains the following components:

`N`	the total sample size
`n0`	sample size in Group 0 (control group)
`n1`	sample size in Group 1 (treatment group)
`two_sided`	logical, `TRUE`, if the estimated sample size refers to a two-sided test
`alpha`	type I error rate used in sample size estimation
`power`	target power used in sample size estimation
`effect`	effect size used in sample size estimation
`effect_type`	short description of the type of effect size
`comment`	additional comment, if there is any
`call`	the matched call.

References

Cundill, B., & Alexander, N. D. E. (2015). Sample size calculations for skewed distributions. *BMC Medical Research Methodology*, 15(1), 1–9. https://doi.org/10.1186/s12874-015-0023-0

Examples

```
n_poisson(mean0 = 5, effect = 0.3)
```

---

pemp *Empirical cumulative density function (ECDF)*

Description

Empirical cumulative density function based on a sample of observations, as used by described by Chakraborti (2006).

Usage

```
pemp(q, sample)
```

Arguments

q	numeric vector of values to evaluate
sample	numeric vector of sample values to base the ECDF on

Value

Returns the probabilities that a value drawn at random from the empirical cumulative density based on *sample* is smaller than or equal to the elements of x.

References

Chakraborti, S., Hong, B., & Van De Wiel, M. A. (2006). A note on sample size determination for a nonparametric test of location. Technometrics, 48(1), 88–94. https://doi.org/10.1198/004017005000000193

Examples

```
x <- 1:5
pemp(1, x)
```

qemp *Empirical quantile function*

Description

Empirical quantile function, i.e. inverse of the empirical cumulative density function `pemp()`. Based on the latter function as presented by Chakraborti (2006).

Usage

```
qemp(p, sample)
```

Arguments

`p`	probability, can be a vector
`sample`	numeric vector of sample values to base the ECDF on

Value

Returns the value for which `pemp(x,sample)` = `p`, i.e. the probability that a value drawn at random from the ECDF is smaller or equal to `x` is `p`.

References

Chakraborti, S., Hong, B., & Van De Wiel, M. A. (2006). A note on sample size determination for a nonparametric test of location. Technometrics, 48(1), 88–94. https://doi.org/10.1198/004017005000000193

Examples

```
x <- 1:5
qemp(0.1, x)
```

remp *Draws random values from the ECDF obtained from* `sample`

Description

Based on the empirical cumulative density function as presented by Chakraborti (2006).

Usage

```
remp(n, sample)
```

Arguments

`n`	integer, number of samples to be drawn
`sample`	numeric vector of sample values to base the ECDF on

Value

numeric vector of random values drawn from the ECDF

References

Chakraborti, S., Hong, B., & Van De Wiel, M. A. (2006). A note on sample size determination for a nonparametric test of location. Technometrics, 48(1), 88–94. https://doi.org/10.1198/004017005000000193

Examples

```
x <- 1:5
remp(10, x)
```

`resample_n_locshift` *Compute a distribution of estimates of N based on two pilot samples.*

Description

Estimation of sample sizes based on resampled pilot samples from the empirical cumulative density. Based on the work of Chakraborti, Hong, & van de Wiel (2006).

Usage

```
resample_n_locshift(
  s1,
  s2,
  delta,
  alpha = 0.05,
  power = 0.9,
  n_resamples = 500,
  q = 0.5
)
```

Arguments

`s1, s2`	Pilot samples
`delta`	numeric value, location shift parameter δ
`alpha`	Type I error probability
`power`	1 - Type II error probability, the desired statistical power
`n_resamples`	number of resamples to use in bootstrapping
`q`	size of group0 relative to total sample size.

Details

WARNING: Note that the estimation has high variability due to its dependence on pilot samples. The smaller the pilot sample, the more uncertain is the estimation of the required sample size. In a simulation study, we found that the method may also be inaccurate on average, depending on the investigated data.

Value

numeric vector of sample size estimates (total sample size)

References

Chakraborti, S., Hong, B., & van de Wiel, M. A. (2006). A note on sample size determination for a nonparametric test of location. Technometrics, 48(1), 88–94. https://doi.org/10.1198/004017005000000193

Index