

Package ‘parasiteR’

July 14, 2026

Type Package

Title A Theoretical-Practical Approach to Parasitological Data Analysis

Description Standardizes and streamlines the processing of parasitological data by integrating descriptive analyses of parasite count distributions, automated calculation of parasitological indices and their dispersion measures, and intuitive visualizations for representing these metrics (Bush et al. 1997 <[doi:10.2307/3284227](https://doi.org/10.2307/3284227)>, Reiczigel et al. 2019 <[doi:10.1016/j.pt.2019.01.003](https://doi.org/10.1016/j.pt.2019.01.003)>).

License GPL (>= 3)

Version 1.1.1

Encoding UTF-8

Imports rlang, ggplot2, magrittr, dplyr, tidyr, stats, BlakerCI, boot, readr, binom, bbmle, readr, MASS

Depends R (>= 4.1.0)

LazyData true

Config/roxygen2/version 8.0.0

RoxygenNote 7.3.3

NeedsCompilation no

Author Exequiel Oscar Furlan [aut] (ORCID: <<https://orcid.org/0000-0002-5642-3956>>),
Juan Manuel Cabrera [aut, cre, cph] (ORCID: <<https://orcid.org/0000-0003-4378-1495>>),
Elisa Helman [aut] (ORCID: <<https://orcid.org/0000-0002-6663-0334>>)

Maintainer Juan Manuel Cabrera <juan.cabrera@uner.edu.ar>

Repository CRAN

Date/Publication 2026-07-13 23:10:02 UTC

Contents

para_abundance_CI	2
para_aggregation_CI	5
para_data	8

para_descriptors	9
para_explo_abund	12
para_explo_prev	14
para_intensity_CI	15
para_plot_CI	18
para_prevalence_CI	20
para_read_data	23

Index	25
--------------	-----------

para_abundance_CI	<i>Mean or median abundance estimation and confidence intervals</i>
-------------------	---

Description

This function calculates point estimates and confidence intervals (CIs) for parasite abundance, using either the mean or the median as a measure of central tendency. These intervals are estimated via a non-parametric bootstrap approach based on resampling (permutations) of the observed data. Specifically, the function implements bias-corrected and accelerated BCa bootstrap intervals, which adjust for both bias and skewness in the bootstrap distribution. This approach does not assume a specific underlying distribution and is particularly robust for overdispersed and zero-inflated parasitological data.

Usage

```
para_abundance_CI(dataset, c_median = TRUE,
  sp_cols, group_vars = NULL, perm = 2000, decimal_places = 2,
  ci_method = "bca", combine_ci = FALSE, conf_level = 0.95,
  save_csv = FALSE, verbose = FALSE)
```

Arguments

dataset	Data frame with parasitological data.
c_median	Logical. If TRUE, the results will include the median as a central tendency of measure; if FALSE, the results will include the mean of the data.
sp_cols	Vector with the names of the columns containing abundance of parasites (taxa) to calculate the parasitological descriptors.
group_vars	Vector with the names of categorical variables used to define groups (e.g., "Sex", "Site"). Default = NULL.
perm	Number of permutations to perform for CIs estimation. Default = 2000.
decimal_places	Number of decimal places to include in the calculation. Default = 2.
ci_method	Method for bootstrap CIs. Options are "bca" (bias-corrected and accelerated) or "perc" (percentile). Default = "bca".
combine_ci	Logical. If TRUE, the interval is expressed as a single column (min - max). If FALSE, the interval is split into separate lower and upper limit columns.
conf_level	Confidence level for the interval estimation (e.g., 0.95 for 95% CIs).

save_csv	Logical. If TRUE, the resulting table is automatically exported as a .CSV file to the current working directory. Default = FALSE.
verbose	Logical. If TRUE, progress messages should be given. Default = FALSE.

Details

Parasite abundance is defined as the number of individuals of a given parasite taxon per host. For each taxon, abundance metrics are calculated based on the observed counts across hosts. The function reshapes the dataset into long format and computes abundance statistics for each parasite taxon and grouping combination (if specified). The following are estimated:

- A is the total parasite abundance
- nH is the number of hosts analyzed
- nH_inf is the number of infected hosts

Depending on the argument `c_median`, the function calculates:

- Mean abundance MeanA: average number of parasites per host
- Median abundance MedA: median number of parasites per host

Confidence intervals are estimated using a non-parametric bootstrap approach. Specifically, bias-corrected and accelerated (BCa) bootstrap intervals are computed by resampling the observed abundance values with replacement a specified number of times `perm`. This method adjusts for both bias and skewness in the bootstrap distribution. Statistical considerations: parasite abundance data are typically overdispersed and zero-inflated, making parametric assumptions inappropriate in many cases. The use of bootstrap methods allows robust estimation of confidence intervals without assuming normality. Mean abundance is sensitive to extreme values, whereas median abundance provides a more robust measure under highly skewed distributions. When sample size is small, bootstrap confidence intervals may be unstable or wide, and results should be interpreted with caution. The interpretation of results remains the responsibility of the user.

Value

A data frame containing abundance estimates and CIs for each parasite taxon, either globally or by group. The following variables are returned:

- nH: Number of hosts analyzed
- nH_inf: Number of infected hosts
- A: Total parasite abundance
- MeanA: Mean parasite abundance
- MedA: Median parasite abundance
- Lower_<ci_method>: Lower bound of the CI estimated using the method specified in `ci_method` ("bca" or "perc").
- Upper_<ci_method>: Upper bound of the CI estimated using the method specified in `ci_method` ("bca" or "perc").
- CI_<ci_method>: If `combine_ci` = TRUE, confidence interval stored as a single column (Lower_<ci_method> - Upper_<ci_method>).

- Observation: Categorical description of the data context:
 - "Not analyzed": No valid observations are available for the given combination (all values are missing or the combination is absent in the dataset); therefore, no estimates can be computed.
 - "One host analyzed": Only a single host analyzed is available for the given combination; thus, no population-level inference is possible and statistical summary measures are not estimated.
 - "No hosts infested": Hosts are present for the given combination, but none are infested (abundance = 0 for all observations); consequently, no statistical summary measures of abundance or intensity can be estimated.
 - "One host infested": Only a single infested host is recorded for the given combination; therefore, no sample-based estimation of intensity or related summary measures is possible.
 - "Multiple hosts infested": More than one infested host is recorded for the given combination, allowing the estimation of summary measures.

Author(s)

Juan Manuel Cabrera, Exequiel Furlan and Elisa Helman

References

- Bush, A.O., Lafferty, K.D., Lotz, J.M., Shostak, A.W. (1997). Parasitology meets ecology on its own terms: Margolis revisited. *Journal of Parasitology*, 83(4), 575–583.
- Reiczigel, J., Marozzi, M., Fabian, I., Rózsa, L. (2019). Biostatistics for parasitologists – a primer to quantitative parasitology. *Trends in Parasitology*, 35(4), 277–281.

Examples

```
#Calculate the CI for the median abundance
med_abun_CI <- para_abundance_CI(para_data$dataset,
                                c_median = TRUE,
                                sp_cols = c("Sp1"),
                                group_vars = c("Site"),
                                decimal_places = 2,
                                ci_method = "bca",
                                conf_level = 0.95,
                                combine_ci = TRUE,
                                save_csv = FALSE,
                                verbose = TRUE)

med_abun_CI

#Calculate the CI for the mean abundance
mean_abun_CI <- para_abundance_CI(para_data$dataset,
                                  c_median = FALSE,
                                  sp_cols = c("Sp1"),
                                  group_vars = c("Site"),
                                  decimal_places = 2,
                                  ci_method = "bca",
                                  conf_level = 0.95,
                                  combine_ci = TRUE,
```

```

                                save_csv = FALSE,
                                verbose = TRUE)

mean_abun_CI

```

para_aggregation_CI	<i>Aggregation indices and confidence intervals for parasite distributions</i>
---------------------	--

Description

This function calculates point estimates and bootstrap confidence intervals (CIs) for six commonly used measures of parasite aggregation among hosts. Following the recommendations of Morrill et al. (2023) and Reiczigel et al. (2019), CIs are estimated via non-parametric bootstrap using the bias-corrected and accelerated (BCa) method, except for the parameter k of the negative binomial distribution, for which maximum likelihood estimation with likelihood profile confidence intervals is used. Parasite aggregation refers to the nearly universal phenomenon where most hosts harbor few or no parasites, while a minority of hosts harbor heavy infections. This pattern has important consequences for host health, parasite transmission, and population dynamics. The six aggregation measures included in this function capture different aspects of this phenomenon and can be grouped into three distinct categories based on their properties and interpretations.

Usage

```

para_aggregation_CI(dataset, sp_cols, group_vars = NULL, perm = 2000,
  decimal_places = 2, ci_method= "bca", combine_ci = FALSE,
  conf_level = 0.95, save_csv = FALSE, verbose = FALSE)

```

Arguments

dataset	Data frame with parasitological data.
sp_cols	Vector with the names of the columns containing abundance of parasites (taxa) to calculate the parasitological descriptors.
group_vars	Vector with the names of categorical variables used to define groups (e.g., "Sex", "Site"). Default is NULL.
perm	Number of permutations to perform for CIs estimation. Default is 2000.
decimal_places	Number of decimal places to include in the calculation. Default is 2.
ci_method	Method for bootstrap CIs. Options are "bca" (bias-corrected and accelerated) or "perc" (percentile). Default = "bca".
combine_ci	Logical. If TRUE, the interval is expressed as a single column (min - max). If FALSE, the interval is split into separate lower and upper limit columns
conf_level	Confidence level for the interval estimation (e.g., 0.95 for 95% CIs).
save_csv	Logical. If TRUE, the resulting table is automatically exported as a .CSV file to the current working directory. Default = FALSE.
verbose	Logical. If TRUE, progress messages should be given. Default = FALSE.

Details

Parasite aggregation is quantified through six indices that capture different attributes of how parasites are distributed among hosts. Following Morrill et al. (2023), these measures can be grouped into three categories based on their mathematical properties and biological interpretations:

- **Abundance-driven measures (sensitive to mean abundance):**

- VMR (Variance-to-Mean Ratio): Calculated as the variance divided by the mean abundance. Values > 1 indicate aggregation, values $= 1$ indicate a random (Poisson) distribution, and values < 1 indicate a more even distribution. This measure is strongly correlated with mean abundance, making it unsuitable for comparing aggregation across samples with different mean abundances.
- Mean crowding: Represents the average number of conspecific parasites sharing a host from the perspective of an individual parasite. Higher values indicate greater crowding experienced by parasites within hosts. This measure is strongly positively correlated with mean abundance, making it unsuitable for comparing aggregation across samples with different mean abundances.

- **Distribution-driven measures (independent of mean abundance):**

- k of the negative binomial distribution: An inverse measure of aggregation where smaller values indicate higher aggregation. When k tends towards infinity, the distribution approaches a Poisson (random) pattern. This measure has a clear biological interpretation in terms of mean crowding: $1/k$ represents the proportion of the mean abundance by which mean crowding exceeds the mean abundance. Importantly, k is not necessarily correlated with mean abundance, contrary to some expectations in the literature.
- Patchiness: Calculated as mean crowding divided by the mean abundance. This measure describes how many times more "crowded" an average parasite is compared to a hypothetical random distribution. Like k , patchiness is not correlated with mean abundance and does not require that the parasite distribution fits a negative binomial.

- **Inequality-driven measures (based on Lorenz curves):**

- Poulin's D: Ranges from 0 (perfectly even distribution) to 1 (maximum possible aggregation, all parasites in one host). This measure is based on the Lorenz curve and is particularly suitable for comparing aggregation across samples or different host-parasite associations. It shows a strong negative correlation with prevalence.
- Hoover's index: Also ranges from 0 to 1 and represents the proportion of parasites that would need to be redistributed to achieve a perfectly even distribution among hosts. For example, a value of 0.7 indicates that 70% of the parasites would need to be moved to achieve evenness. This measure has a clear biological interpretation but can be constrained at low prevalence values (often equaling $1 - \text{prevalence}$ when all infected hosts have infrapopulations at least as large as the mean abundance).

Confidence intervals are estimated using a non-parametric bootstrap approach. Specifically, BCa bootstrap intervals are computed by resampling the observed abundance values with replacement a specified number of times (perm). This method adjusts for both bias and skewness in the bootstrap distribution and is recommended for aggregation measures by Morrill et al. (2023) and Reiczigel et al. (2019). For the parameter k of the negative binomial distribution, confidence intervals are estimated using maximum likelihood estimation with likelihood profile confidence intervals, following the specific recommendation of Morrill et al. (2023). For all other indices (VMR, mean crowding,

patchiness, Poulin's D, and Hoover's index), BCa bootstrap intervals are used. Statistical considerations: Parasite abundance data are typically overdispersed and zero-inflated, making parametric assumptions inappropriate in many cases. The use of bootstrap methods allows robust estimation of confidence intervals without assuming normality. When sample size is small (fewer than 3 hosts), CIs may not be estimable and results should be interpreted with caution. The interpretation of results remains the responsibility of the user.

Value

A data frame containing aggregation estimates and CIs for each parasite taxon, either globally or by group. The following variables are returned:

- nH: Number of hosts analyzed
- nH_inf: Number of infected hosts
- VMR: Variance-to-Mean Ratio
- mean_crowding: Mean crowding index
- patchiness: Patchiness index
- poulin_D: Poulin's D index
- hoover: Hoover's index
- k: Parameter of the negative binomial distribution
- Lower_<ci_method>: Lower bound of the CI estimated using the method specified in ci_method ("bca" or "perc").
- Upper_<ci_method>: Upper bound of the CI estimated using the method specified in ci_method ("bca" or "perc").
- CI_<ci_method>: If combine_ci = TRUE, confidence interval stored as a single column (Lower_<ci_method> - Upper_<ci_method>).
- Observation: Categorical description of the data context:
 - "Not analyzed": No valid observations are available for the given combination (all values are missing or the combination is absent in the dataset); therefore, no estimates can be computed.
 - "One host analyzed": Only a single host analyzed is available for the given combination; thus, no population-level inference is possible and statistical summary measures are not estimated.
 - "No hosts infested": Hosts are present for the given combination, but none are infested (abundance = 0 for all observations); consequently, no statistical summary measures of abundance or intensity can be estimated.
 - "One host infested": Only a single infested host is recorded for the given combination; therefore, no sample-based estimation of intensity or related summary measures is possible.
 - "Multiple hosts infested": More than one infested host is recorded for the given combination, allowing the estimation of summary measures.

Author(s)

Juan Manuel Cabrera, Exequiel Furlan and Elisa Helman

References

- Bush, A.O., Lafferty, K.D., Lotz, J.M., Shostak, A.W. (1997). Parasitology meets ecology on its own terms: Margolis revisited. *Journal of Parasitology*, 83(4), 575–583.
- Reiczig, J., Marozzi, M., Fabian, I., Rózsa, L. (2019). Biostatistics for parasitologists – a primer to quantitative parasitology. *Trends in Parasitology*, 35(4), 277–281.
- Morrill, A., Poulin, R., Forbes, M., 2023. Interrelationships and properties of parasite aggregation measures: a user’s guide. *Int. Journal of Parasitology* 53, 763–776.

Examples

```
# Calculate aggregation indices with BCa confidence intervals
agg_CI <- para_aggregation_CI(para_data$dataset,
                             sp_cols = c("Sp1", "Sp2"),
                             group_vars = c("Site", "Sp_host"),
                             perm = 2000,
                             decimal_places = 2,
                             ci_method = "bca",
                             combine_ci = FALSE,
                             conf_level = 0.95,
                             save_csv = FALSE,
                             verbose = TRUE)

agg_CI
```

para_data	<i>Simulated parasite abundance data for multiple species across hosts and sites</i>
-----------	--

Description

This dataset contains hypothetical generated parasite count data representing multiple parasite species infecting individual hosts across different sampling sites. Each row corresponds to a single sampling unit (i.e., an individual host), and parasite abundance is recorded as counts for each parasite species (Sp1–Sp4).

Usage

```
para_data
```

Format

'para_data' A list with 4 elements

- dataset A data frame with 81 rows and 6 columns:
 - Site: Factor or character. Sampling location where hosts were collected (sites A, B, C and D). Multiple hosts can belong to the same site.

- Factor or character. Host species identifier. In this dataset, each site includes up to two host species (HostA, HostB), although some site–host combinations may be absent by design.
 - Sp1: Integer. Abundance (count) of parasite species 1 per host. Simulated using an aggregated (negative binomial) distribution across all sites.
 - Sp2: Integer. Abundance of parasite species 2 per host. Present only in Sites A and B; missing (NA) in Site C to represent non-analyzed combinations.
 - Sp3: Integer. Abundance of parasite species 3 per host. Designed to represent heterogeneous infection patterns: full infection in one host group, rare infection in another and absence elsewhere.
 - Sp4: Integer. Integer. Abundance of parasite species 4 per host. Includes several edge cases: only one host examined, no infected hosts, a single infected host, multiple infected hosts.
- factors_v: A list of columns with factor values.
 - num_v: A list of columns with numeric values.
 - summ: A summary of the loaded data. Check [summary](#).

Details

The dataset was intentionally constructed to reproduce common scenarios encountered in parasitological studies, rather than to reflect a specific empirical system. These scenarios include:

- zero-inflated parasite distributions
- aggregated parasite abundances
- missing data (non-analyzed host–parasite combinations)
- rare infections (single infected host)
- absence of infection
- small sample sizes for specific host–site combinations

This structure allows testing and demonstrating the behavior of analytical functions under realistic and edge-case conditions.

para_descriptors

Parasitological descriptors and summary statistics

Description

Computes standard parasitological descriptors and classical summary statistics from parasite abundance data, optionally stratified by grouping variables.

Usage

```
para_descriptors(dataset, sp_cols = NULL, group_vars = NULL,
  decimal_places = 2, save_csv = FALSE, verbose = FALSE)
```

Arguments

dataset	Data frame with parasitic abundance data.
sp_cols	Vector with the names or indices of the species columns.
group_vars	Vector with the names of the categorical variables to consider (e.g., 'Sex', 'Site').
decimal_places	Number of decimal places to round the values.
save_csv	Logical. If TRUE, the resulting table is automatically exported as a .CSV file to the current working directory. Default = FALSE.
verbose	Logical. If TRUE, progress messages should be given. Default = FALSE.

Details

The `para_descriptors` function provides a practical and efficient way to estimate the main parasitological descriptors commonly used in ecological and parasitological studies. Calculations can be performed globally or at different hierarchical levels defined by grouping variables.

The function computes descriptors based on parasite abundance per sampling unit (e.g., host, site, or pooled hosts), following standard definitions:

- Prevalence (P): Proportion of infected hosts.
- Abundance (A): Total number of parasites recorded.
- Intensity (I): Number of parasites per infected host.

Statistical validity and sample size considerations: The estimation of summary statistics is subject to fundamental statistical constraints related to sample size and variability.

- **Host population (nH):** Summary statistics for abundance (e.g., mean, median, standard deviation) are only meaningful when more than one host is analyzed ($nH > 1$). When only a single host is available ($nH = 1$), no population of hosts exists and therefore no variability can be estimated. In such cases, the observed value is reported, but it should not be interpreted as a population-level descriptor.
- **Infected host population (nH_inf):** Similarly, intensity-based descriptors are only meaningful when more than one infected host is available ($nH_inf > 1$). When only one infected host is present, the parasite count corresponds to a single observation rather than a distribution and summary statistics of intensity cannot be formally estimated.

These constraints reflect a fundamental principle: statistical descriptors require variability, and variability requires more than one observational unit. When this condition is not met, results should be interpreted cautiously, and no generalization beyond the observed case is justified. Handling of special cases: The function automatically adjusts calculations depending on data availability:

- When no data are available → results are reported as NA.
- When hosts are analyzed but none are infected → prevalence is 0 and intensity measures are not computed.
- When only one host or one infected host is available → corresponding summary statistics are not computed, and interpretation should be limited to the observed value.

The selection and interpretation of descriptors remain the responsibility of the user, particularly when working with small sample sizes.

Value

A data frame containing the calculated parasitological descriptors for each parasite taxon, either globally or by group (if grouping variables are specified). The following variables are returned:

- nH: Number of hosts analyzed
- nH_inf: Number of infected hosts
- A: Total parasite abundance
- min: Minimum parasite count
- max: Maximum parasite count
- P: Parasitic prevalence
- MeanA: Mean parasitic abundance
- MeanA_sd: Standard deviation of mean parasite abundance
- A_iqr: Interquartile range of mean parasite abundance
- MedA: Median parasite abundance
- MedA_sd: Median absolute deviation of parasite abundance
- MeanI: Mean parasite intensity
- MeanI_sd: Standard deviation of mean parasite intensity
- I_iqr: Interquartile range of mean parasite intensity
- MedI: Median parasite intensity
- MedI_sd: Median absolute deviation of parasite intensity
- Observation: Qualitative descriptor indicating data availability and sample structure for each hierarchical combination:
 - "Not analyzed": No valid observations are available for the given combination (all values are missing or the combination is absent in the dataset); therefore, no estimates can be computed.
 - "One host analyzed": Only a single host analyzed is available for the given combination; thus, no population-level inference is possible and statical summary measures are not estimated.
 - "No hosts infested": Hosts are present for the given combination, but none are infested (abundance = 0 for all observations); consequently, no statical summary measures of abundance or intensity can be estimated.
 - "One host infested": Only a single infested host is recorded for the given combination; therefore, no sample-based estimation of intensity or related summary measures is possible.
 - "Multiple hosts infested": More than one infested host is recorded for the given combination, allowing the estimation of summary measures.

Author(s)

Juan Manuel Cabrera, Exequiel Furlan and Elisa Helman

References

- Bush, A.O., Lafferty, K.D., Lotz, J.M., Shostak, A.W. (1997). Parasitology meets ecology on its own terms: Margolis revisited. *Journal of Parasitology*, 83(4), 575–583.
- Reiczigel, J., Marozzi, M., Fabian, I., Rózsa, L. (2019). Biostatistics for parasitologists – a primer to quantitative parasitology. *Trends in Parasitology*, 35(4), 277–281.

Examples

```
p_descriptor <- para_descriptors(para_data$dataset,
                                sp_cols = c("Sp1", "Sp2", "Sp3", "Sp4"),
                                group_vars = c("Site", "Sp_host"),
                                decimal_places = 2,
                                save_csv = FALSE,
                                verbose = FALSE)

p_descriptor
```

para_explo_abund	<i>Exploratory plots of parasite abundance distributions</i>
------------------	--

Description

Generates exploratory visualizations of parasite abundance distributions across taxa and optional grouping variables. The function produces histograms combined with kernel density curves to facilitate the assessment of distributional patterns, including skewness, dispersion and zero inflation.

Usage

```
para_explo_abund(dataset, sp_cols, group_vars = NULL,
  bins = 30, n_col = NULL, save_fig = "none", verbose = FALSE)
```

Arguments

dataset	Data frame containing parasite data.
sp_cols	Vector with the names of the columns containing parasite abundance (taxa) to be plotted.
group_vars	Vector with the names of categorical variables used to define groups (e.g., "Sex", "Site"). Default = NULL.
bins	Integer specifying the number of bins used in the histogram. Higher values provide finer resolution but may introduce noise, while lower values produce smoother but less detailed distributions. Default = 30.
n_col	Integer specifying the number of columns in the faceted plot layout. If NULL, the number of columns is determined automatically by ggplot2. Default = NULL.
save_fig	Character string indicating the file format used to export the plot in the current working directory. Valid options are "none" (the plot is not saved), "pdf", "jpeg", "tiff", and "svg".
verbose	Logical. If TRUE, progress messages should be given. Default = FALSE.

Details

The function reshapes the input dataset into a long format, where parasite taxa are treated as a single variable and their abundances as observations. For each parasite taxon and combination of grouping variables (if provided), the function generates:

- A **histogram** representing the distribution of parasite abundance values
- A **kernel density curve** (when sufficient data are available), providing a smoothed approximation of the underlying distribution.

Both elements are scaled to represent density, allowing direct comparison between distributions. These plots are intended for exploratory purposes and should not be used as formal inference tools. Faceting is applied to display each taxon and grouping combination in separate panels. Special cases are handled as follows:

- When all abundance values for a given combination are zero, no histogram or density curve is drawn and a message is displayed indicating that the parasite was not recorded for that combination.
- When the number of observations is insufficient (less than 2), a message is displayed indicating that there is not enough data to compute a meaningful distribution.
- Density curves are only computed when there are more than two observations and non-zero variation.

All plots use independent scales (free scales) to better represent the variability within each facet.

Value

A ggplot2 object containing the generated faceted plots. This object can be further customized using standard ggplot2 functions.

Author(s)

Juan Manuel Cabrera, Exequiel Furlan and Elisa Helman

Examples

```
#Species 1 and 2
```

```
para_explo_abund (para_data$dataset,
  sp_cols = c("Sp1", "Sp2"),
  group_vars = c("Site", "Sp_host"),
  bins = 30,
  n_col = 4,
  save_fig = "none",
  verbose = TRUE)
```

```
#Species 3 and 4
```

```
para_explo_abund (para_data$dataset,
  sp_cols = c("Sp3", "Sp4"),
  group_vars = c("Site", "Sp_host"),
```

```
bins = 30,
n_col = 4,
save_fig = "none",
verbose = TRUE)
```

para_explo_prev

*Exploratory plots of parasite prevalence***Description**

Generates exploratory visualizations of parasite prevalence across taxa and optional grouping variables. The function produces stacked bar plots showing the proportion of infested and non-infested hosts, facilitating the assessment of prevalence patterns across hierarchical combinations.

Usage

```
para_explo_prev(dataset, sp_cols, group_vars = NULL,
  n_col = NULL, save_fig = "none", verbose = FALSE)
```

Arguments

dataset	Data frame containing parasite data.
sp_cols	Vector with the names of the columns containing parasite abundance (taxa) to be plotted.
group_vars	Vector with the names of categorical variables used to define groups (e.g., "Sex", "Site"). Default = NULL.
n_col	Integer specifying the number of columns in the faceted plot layout. If NULL, the number of columns is determined automatically by ggplot2. Default = NULL.
save_fig	Character string indicating the file format used to export the plot in the current working directory. Valid options are "none" (the plot is not saved), "pdf", "jpeg", "tiff", and "svg".
verbose	Logical. If TRUE, progress messages should be given. Default = FALSE.

Details

The function reshapes the dataset into long format and calculates prevalence as the proportion of infested hosts (hosts with parasite counts > 0) relative to the number of analyzed hosts for each parasite taxon and grouping combination. For each combination, the function generates:

- The proportion of infested hosts.
- The proportion of non-infested hosts.

Faceting is applied to display each parasite taxon and grouping combination in separate panels. Special cases are handled as follows:

- When no observations are available (all values are missing or the combination is absent), a message is displayed indicating that the data were not analyzed.
- When only one host is available, a message is displayed indicating that the sample size is insufficient for prevalence estimation.
- When all observed values are zero, a message is displayed indicating that the parasite was not recorded for that combination.

All proportions are expressed on a 0–1 scale. These plots are intended for exploratory purposes and should not be used as formal inference tools.

Value

A ggplot2 object containing the generated faceted stacked bar plots. This object can be further customized using standard ggplot2 functions.

Author(s)

Juan Manuel Cabrera, Exequiel Furlan and Elisa Helman

Examples

#Species 1 and 2

```
para_explo_prev(para_data$dataset,
  sp_cols = c("Sp1", "Sp2"),
  group_vars = c("Site", "Sp_host"),
  n_col = 4,
  save_fig = "none",
  verbose = TRUE)
```

#Species 3 and 4

```
para_explo_prev(para_data$dataset,
  sp_cols = c("Sp3", "Sp4"),
  group_vars = c("Site", "Sp_host"),
  n_col = 4,
  save_fig = "none",
  verbose = TRUE)
```

Description

This function calculates point estimates and confidence intervals (CIs) for parasite intensity, using either the mean or the median as a measure of central tendency. CIs are estimated via a non-parametric bootstrap approach based on resampling (permutations) of the observed data. Specifically, the function implements bias-corrected and accelerated (BCa) bootstrap intervals, which adjust for both bias and skewness in the bootstrap distribution. This approach does not assume a specific underlying distribution and is particularly robust for overdispersed and zero-inflated parasitological data.

Usage

```
para_intensity_CI(dataset, c_median = TRUE, sp_cols, group_vars = NULL,
  perm = 2000, decimal_places = 2, ci_method= "bca", combine_ci = FALSE,
  conf_level = 0.95, save_csv = FALSE, verbose = FALSE)
```

Arguments

dataset	Data frame with parasitological data.
c_median	Logical. If TRUE, the results will include the median as a central tendency of measure; if FALSE, the results will include the mean of the data.
sp_cols	Vector with the names of the columns containing abundance of parasites (taxa) to calculate the parasitological descriptors.
group_vars	Vector with the names of categorical variables used to define groups (e.g., "Sex", "Site"). Default is NULL.
perm	Number of permutations to perform for CIs estimation. Default is 2000.
decimal_places	Number of decimal places to include in the calculation. Default is 2.
ci_method	Method for bootstrap CIs. Options are "bca" (bias-corrected and accelerated) or "perc" (percentile). Default = "bca".
combine_ci	Logical. If TRUE, the interval is expressed as a single column (min - max). If FALSE, the interval is split into separate lower and upper limit columns
conf_level	Confidence level for the interval estimation (e.g., 0.95 for 95% CIs).
save_csv	Logical. If TRUE, the resulting table is automatically exported as a .CSV file to the current working directory. Default = FALSE.
verbose	Logical. If TRUE, progress messages should be given. Default = FALSE.

Details

Parasite intensity is defined as the number of individuals of a given parasite taxon per infested host. For each taxon, intensity metrics are calculated based only on hosts with parasite counts greater than zero. The function reshapes the dataset into long format and computes intensity statistics for each parasite taxon and grouping combination (if specified). The following are estimated:

- A is the total parasite abundance
- nH is the number of hosts analyzed
- nH_inf is the number of infected hosts

Depending on the argument `c_median`, the function calculates:

- Mean intensity `MeanI`: average number of parasites per infested host
- Median intensity `MedI`: median number of parasites per infested host

CIs are estimated using a non-parametric bootstrap approach. Specifically, BCa bootstrap intervals are computed by resampling the observed intensity values with replacement a specified number of times `perm`. This method adjusts for both bias and skewness in the bootstrap distribution. Statistical considerations: parasite intensity data are typically right-skewed and may exhibit high variability due to aggregation among hosts. The use of bootstrap methods allows robust estimation of confidence intervals without assuming normality. Mean intensity is sensitive to extreme values, whereas median intensity provides a more robust measure under highly skewed distributions. When the number of infested hosts is small, bootstrap CIs intervals may be unstable or wide, and results should be interpreted with caution. The interpretation of results remains the responsibility of the user.

Value

A data frame containing abundance estimates and CIs for each parasite taxon, either globally or by group. The following variables are returned:

- `nH`: Number of hosts analyzed
- `nH_inf`: Number of infected hosts
- `A`: Total parasite abundance
- `MeanA`: Mean parasite intensity
- `MedA`: Median parasite intensity
- `Lower_<ci_method>`: Lower bound of the CI estimated using the method specified in `ci_method` ("bca" or "perc").
- `Upper_<ci_method>`: Upper bound of the CI estimated using the method specified in `ci_method` ("bca" or "perc").
- `CI_<ci_method>`: If `combine_ci = TRUE`, CI stored as a single column (`Lower_<ci_method>` - `Upper_<ci_method>`).
- `Observation`: Categorical description of the data context:
 - "Not analyzed": No valid observations are available for the given combination (all values are missing or the combination is absent in the dataset); therefore, no estimates can be computed.
 - "One host analyzed": Only a single host analyzed is available for the given combination; thus, no population-level inference is possible and statistical summary measures are not estimated.
 - "No hosts infested": Hosts are present for the given combination, but none are infested (abundance = 0 for all observations); consequently, no statistical summary measures of abundance or intensity can be estimated.
 - "One host infested": Only a single infested host is recorded for the given combination; therefore, no sample-based estimation of intensity or related summary measures is possible.
 - "Multiple hosts infested": More than one infested host is recorded for the given combination, allowing the estimation of summary measures.

Author(s)

Juan Manuel Cabrera, Exequiel Furlan and Elisa Helman

References

Bush, A.O., Lafferty, K.D., Lotz, J.M., Shostak, A.W. (1997). Parasitology meets ecology on its own terms: Margolis revisited. *Journal of Parasitology*, 83(4), 575–583.

Reiczigel, J., Marozzi, M., Fabian, I., Rózsa, L. (2019). Biostatistics for parasitologists – a primer to quantitative parasitology. *Trends in Parasitology*, 35(4), 277–281.

Examples

```
# Calculate of the CI for the median intensity
med_int_CI <- para_intensity_CI(para_data$dataset,
                               c_median = TRUE,
                               sp_cols = c("Sp1"),
                               group_vars = c("Site"),
                               decimal_places = 2,
                               ci_method = "bca",
                               conf_level = 0.95,
                               combine_ci = TRUE,
                               save_csv = FALSE,
                               verbose = TRUE)

med_int_CI

mean_int_CI <- para_intensity_CI(para_data$dataset,
                                c_median = FALSE,
                                sp_cols = c("Sp1"),
                                group_vars = c("Site"),
                                decimal_places = 2,
                                ci_method = "bca",
                                conf_level = 0.95,
                                combine_ci = TRUE,
                                save_csv = FALSE,
                                verbose = TRUE)

mean_int_CI
```

para_plot_CI

Visualization of parasitological descriptor with confidence intervals

Description

This function generates graphical representations of parasitological estimates (abundance, intensity, or prevalence) including their associated confidence intervals. It supports multiple input formats and automatically detects the response variable and CI structure. The function allows flexible grouping, species filtering, and visualization either as faceted plots or separate panels. The input must be the direct, unmodified output returned by one of the compatible CI functions (e.g.,

[para_abundance_CI](#), [para_intensity_CI](#), [para_prevalence_CI](#)) to ensure proper automatic detection of CIs. Manually renamed, summarized, reshaped, or otherwise restructured tables are not supported, as the function relies on the original descriptor and CI columns to automatically identify the estimates and plot the associated uncertainty. It automatically detects:

Usage

```
para_plot_CI(para_df, group_vars, sp_cols = NULL, descriptor = NULL,
  point_color = "blue", line_size = 1, point_size = 3, n_cols = 1,
  include_zeros = TRUE, separate_plots = FALSE, save_fig = "none",
  verbose = FALSE)
```

Arguments

para_df	Data frame returned directly and without modification of the following functions: para_abundance_CI , para_intensity_CI , para_prevalence_CI or para_aggregation_CI .
group_vars	Character vector specifying the variable(s) to be used on the x-axis. Multiple variables will be combined.
sp_cols	Optional vector of parasite taxa to include in the plot. Default is NULL (all taxa are included).
descriptor	Name of the variable to be plotted on the y-axis. If NULL, the function automatically detects a suitable variable (e.g., prevalence, MeanA, MedA, MeanI, MedI).
point_color	Color of the points. Default is "blue".
line_size	Line width of the CI bars. Default is 1.
point_size	Size of the points. Default is 3.
n_cols	Number of columns used in faceted plots. Default is 1.
include_zeros	Logical. If FALSE, zero values are excluded from the plot. Default is TRUE.
separate_plots	Logical. If TRUE, returns a list of plots (one per species). If FALSE, produces a faceted plot. Default is FALSE.
save_fig	Character string indicating the file format used to export the plot in the current working directory. Valid options are "none" (the plot is not saved), "pdf", "jpeg", "tiff", and "svg".
verbose	Logical. If TRUE, progress messages should be given. Default = FALSE.

Details

- The response variable to be plotted.
- The structure of CIs, including:
 - Separate columns (Lower_CI, Upper_CI)
 - Method-specific intervals (e.g., exact or Blaker)
 - Combined intervals stored as a single character column (e.g., "min - max")

Aggregation outputs contain multiple descriptors (e.g., VMR, mean crowding, patchiness, Poulin's D, Hoover index, and k). If descriptor is not specified, `para_plot_CI()` automatically plots the variance-to-mean ratio (VMR). To visualize any other aggregation index, the corresponding descriptor must be explicitly provided (e.g., `descriptor = "poulin_D"`).

When multiple grouping variables are provided in `x_var`, they are combined into a single factor for visualization. CIs are displayed as vertical error bars, and point estimates are overlaid. When multiple parasite taxa are present, results are displayed using faceting or as separate plots. Interpretation of graphical outputs remains the responsibility of the user.

Value

A ggplot object or a list of ggplot objects representing the estimated values and their CIs.

Author(s)

Juan Manuel Cabrera, Exequiel Furlan and Elisa Helman

References

Bush, A.O., Lafferty, K.D., Lotz, J.M., Shostak, A.W. (1997). Parasitology meets ecology on its own terms: Margolis revisited. *Journal of Parasitology*, 83(4), 575–583.

Reiczigel, J., Marozzi, M., Fabian, I., Rózsa, L. (2019). Biostatistics for parasitologists – a primer to quantitative parasitology. *Trends in Parasitology*, 35(4), 277–281.

para_prevalence_CI	<i>Parasite prevalence estimation and confidence intervals</i>
--------------------	--

Description

Estimates parasite prevalence and corresponding confidence intervals from parasite abundance data, optionally stratified by grouping variables. Two types of CIs are provided: exact binomial intervals and Blaker intervals, allowing robust inference across a wide range of sample sizes and prevalence values.

Usage

```
para_prevalence_CI(dataset, sp_cols, group_vars = NULL, decimal_places = 2,
  conf_level = 0.95, ci_method = "wilson", output_type = "proportion",
  combine_ci = FALSE, save_csv = FALSE, verbose = FALSE)
```

Arguments

<code>dataset</code>	Data frame with parasitological data.
<code>sp_cols</code>	Vector with the names of the columns containing abundance of parasites (taxa) to calculate the parasitological descriptors.
<code>group_vars</code>	Vector with the names of categorical variables used to define groups (e.g., "Sex", "Site"). Default is NULL.
<code>decimal_places</code>	Number of decimal places to round the values. Default is 2.
<code>conf_level</code>	Confidence level for the interval estimation (e.g., 0.95 for 95% CIs).

ci_method	Method used to estimate prevalence CIs. Options are "wilson" (Wilson score interval), "exact" (Clopper–Pearson exact interval), or "blaker" (Blaker exact interval). Default is "wilson".
output_type	Format of the result: either "proportion" or "percentage". Default is "proportion".
combine_ci	Logical. If TRUE, the interval is expressed as a single column (min - max). If FALSE, the interval is split into separate lower and upper limit columns.
save_csv	Logical. If TRUE, the resulting table is automatically exported as a .CSV file to the current working directory. Default = FALSE.
verbose	Logical. If TRUE, progress messages should be given. Default = FALSE.

Details

Prevalence is defined as the proportion of hosts infected with a given parasite taxon:

$$P = \frac{nH_{inf}}{nH}$$

where:

- nH is the number of hosts analyzed (non-missing observations)
- nH_{inf} is the number of infected hosts (abundance > 0)

The function reshapes the dataset into long format and computes prevalence for each parasite taxon and grouping combination (if specified). Two types of CIs are calculated:

- **Exact (Clopper–Pearson) interval:** This is an exact binomial CI, conservative but valid for all sample sizes, especially small samples or extreme prevalence values.
- **Blaker interval:** This interval is also exact but less conservative than Clopper–Pearson, providing shorter intervals while maintaining correct coverage.

Statistical considerations:

- Prevalence is a binomial proportion and can be estimated even for small sample sizes.
- However, when sample size is very small (e.g., $nH = 1$), the estimate lacks precision and CIs become uninformative.
- When no infected hosts are observed ($nH_{inf} = 0$), prevalence is 0, and CIs reflect uncertainty around zero.

The interpretation of results, particularly under small sample sizes, remains the responsibility of the user.

Value

A data frame containing prevalence estimates and CIs for each parasite taxon, either globally or by group. The following variables are returned:

- nH : Number of hosts analyzed
- nH_{inf} : Number of infested hosts

para_read_data	<i>Read parasite data</i>
----------------	---------------------------

Description

Load data from a .CSV file

Usage

```
para_read_data(file_name, verbose = FALSE)
```

Arguments

file_name	Name of .CSV table file.
verbose	A logical value indicating if progress messages should be given.

Details

This package includes a specific function to import tables (.CSV files) into the R environment. Each row in the table should correspond to an individual host that was analyzed, while the columns may contain both quantitative and qualitative variables. Columns may represent two principal categories of variables:

- "Host-related variables": Encompassing metadata such as the site of specimen collection, host species, morphophysiological traits, applied experimental treatments, and other relevant descriptors.
- "Parasite-related variables": Denoting parasite abundance per host, typically structured across multiple columns corresponding to the finest available taxonomic resolution (e.g., species, genus, family, order).

Parasite abundance values must be encoded as non-negative integers. It is critical to distinguish between the following:

- 0: Represents a confirmed absence of the parasite in the host specimen.
- NA: Indicates that parasite detection or quantification was not feasible due to methodological or technical limitations.

Value

The function returns:

dataset	A table that can be used as input for other parasiteR functions.
factors_v	A list of columns with factor values.
num_v	A list of columns with numeric values.
summ	A summary of the loaded data. Check <code>summary()</code> function

Author(s)

Juan Manuel Cabrera, Exequiel Furlan and Elisa Helman

Index

* **datasets**

para_data, [8](#)

para_abundance_CI, [2](#), [19](#)

para_aggregation_CI, [5](#), [19](#)

para_data, [8](#)

para_descriptors, [9](#)

para_explo_abund, [12](#)

para_explo_prev, [14](#)

para_intensity_CI, [15](#), [19](#)

para_plot_CI, [18](#)

para_prevalence_CI, [19](#), [20](#)

para_read_data, [23](#)

summary, [9](#)