Package ‘pempi’

October 9, 2023

Type Package

Title Proportion Estimation with Marginal Proxy Information

Version 1.0.0

Date 2023-09-15

LazyData true

Maintainer Stéphane Guerrier <stef.guerrier@gmail.com>

Description A system contains easy-to-use tools for the conditional estimation of the prevalence of an emerging or rare infectious diseases using the methods proposed in Guerrier et al. (2023) <arXiv:2012.10745>.

Depends R (>= 4.0.0)

License AGPL-3

RoxygenNote 7.2.3

Encoding UTF-8

Suggests knitr, rmarkdown

VignetteBuilder knitr

URL https://github.com/stephaneguerrier/pempi

BugReports https://github.com/stephaneguerrier/pempi/issues

NeedsCompilation no

Author Stéphane Guerrier [aut, cre],
Maria-Pia Victoria-Feser [aut],
Christoph Kuzmics [aut]

Repository CRAN

Date/Publication 2023-10-09 12:20:02 UTC

R topics documented:

conditional_mle *Compute MLE based on the full information R1, R2, R3 and R4.*

Description

Proportion estimated using the MLE and confidence intervals based the asymptotic distribution of the estimator.

Usage

```
conditional_mle(
  R1 = NULL,
  R2 = NULL,
  R3 = NULL,
  R4 = NULL,
  n = R1 + R2 + R3 + R4,
  pi0,
  gamma = 0.05,
  alpha0 = 0,
  alpha = 0,
  beta = 0,
  V = NULL,
  ...
)
```

Arguments

`R1` A `numeric` that provides the number of participants in the survey sample that were tested positive with both (medical) testing devices (and are, thus, members of the sub-population).

`R2` A `numeric` that provides the number of participants in the survey sample that are tested positive only with the first testing device (and are, thus, members of the sub-population).

`R3` A `numeric` that provides the number of participants in the survey sample that are tested positive only with the second testing device.

`R4` A `numeric` that provides the number of participants that are tested negative with the second testing device (and are either members of the sub-population and have tested negative with the first testing device or are not members of the sub-population).

`n` A `numeric` that provides the sample size. Default value R1 + R2 + R3 + R4. If this value is provided it is used to verify that R1 + R2 + R3 + R4 = n.

`pi0` A `numeric` that provides the prevalence or proportion of people (in the whole population) who are positive, as measured through a non-random, but systematic sampling (e.g. based on medical selection).

`gamma` A `numeric` that is used to compute a (1 - gamma) confidence region for the proportion. Default value is `0.05`.

`alpha0` A `numeric` that corresponds to the probability that a random participant has been incorrectly declared positive through the nontransparent procedure. In most applications, this probability is likely very close to zero. Default value is `0`.

`alpha` A `numeric` that provides the False Negative (FN) rate for the sample R. Default value is `0`.

`beta` A `numeric` that provides the False Positive (FP) rate for the sample R. Default value is `0`.

`V` A `numeric` that corresponds to the average of squared sampling weights. Default value is `NULL`.

`...` Additional arguments.

Value

A `cpreval` object with the structure:

- estimate: Estimated proportion.
- sd: Estimated standard error of the estimator.
- ci_asym: Asymptotic confidence interval at the 1 - gamma confidence level.
- gamma: Confidence level (i.e. 1 - gamma) for confidence intervals.
- method: Estimation method (in this case mle).
- measurement: A vector with (alpha0, alpha, beta).
- beta0: Estimated false negative rate of the official procedure.
- ci_beta0: Asymptotic confidence interval (1 - gamma confidence level) for beta0.
- boundary: A boolean variable indicating if the estimates falls at the boundary of the parameter space.
- pi0: Value of pi0 (input value).
- sampling: Type of sampling considered ("random" or "weighted").
- V: Average sum of squared sampling weights if weighted/stratified is used (otherwise NULL).
- n: Sample size.
- avar_beta0: Estimated asymptotic variance of beta0
- ...: Additional parameters.

Author(s)

Stephane Guerrier, Maria-Pia Victoria-Feser, Christoph Kuzmics

Examples

```
# Samples without measurement error
X = sim_Rs(theta = 3/100, pi0 = 1/100, n = 1500, seed = 18)
conditional_mle(R1 = X$R1, R2 = X$R2, R3 = X$R3, R4 = X$R4, pi0 = X$pi0)

# With measurement error
X = sim_Rs(theta = 30/1000, pi0 = 10/1000, n = 1500, alpha0 = 0.001,
alpha = 0.01, beta0 = 0.05, beta = 0.05, seed = 18)
conditional_mle(R1 = X$R1, R2 = X$R2, R3 = X$R3, R4 = X$R4, pi0 = X$pi0)
conditional_mle(R1 = X$R1, R2 = X$R2, R3 = X$R3, R4 = X$R4, pi0 = X$pi0,
alpha0 = 0.001, alpha = 0.01, beta = 0.05)
```

`covid19_austria` *COVID-19 Data from Statistics Austria*

Description

Data collected in Austria in 2020 (see e.g. SORA, 2020; Kowarik et al., 2021, for more details), allowing to estimate COVID-19 prevalence.

Usage

```
covid19_austria
```

Format

A `matrix` with 2290 rows and 3 variables:

Y Binary variable, 1 if participant i is tested positive in the survey sample, 0 otherwise.

Z Binary variable, 1 if participant i was declared positive with the official procedure, 0 otherwise.

weights Sampling weights.

Source

Statistics Austria. 2020. "Prävalenz von SARS-CoV-2-Infektionen liegt bei 0.031."

get_prob *Compute sucess probabilities (tau_j's)*

Description

Compute joint probabilities of P(W = j, Y = k) for j, k = 0, 1.

Usage

```
get_prob(theta, pi0, alpha, beta, alpha0)
```

Arguments

`theta` A `numeric` that provides the true prevalence of a given disease.

`pi0` A `numeric` that provides the prevalence or proportion of people (in the whole population) who are positive, as measured through a non-random, but systematic sampling (e.g. based on medical selection).

`alpha` A `numeric` that provides the False Negative (FN) rate for the sample R.

`beta` A `numeric` that provides the False Positive (FP) rate for the sample R.

`alpha0` A `numeric` that corresponds to the probability that a random participant has been incorrectly declared positive through the nontransparent procedure. In most applications, this probability is likely very close to zero.

Value

A `vector` containing tau1, tau2, tau3 and tau4.

Author(s)

Stephane Guerrier

Examples

```
prob1 = get_prob(theta = 0.02, pi0 = 0.01, alpha = 0, beta = 0, alpha0 = 0)
prob1
sum(prob1)

prob2 = get_prob(theta = 0.02, pi0 = 0.01, alpha = 0.001, beta = 0, alpha0 = 0.001)
prob2
sum(prob2)
```

marginal_mle — *Compute (marginalized) MLE based on the partial information R1 and R3.*

Description

Proportion estimated using the MLE and confidence intervals based the asymptotic distribution of the estimator.

Usage

```
marginal_mle(
  R1,
  R3,
  n,
  pi0,
  gamma = 0.05,
  alpha = 0,
  beta = 0,
  alpha0 = 0,
  V = NULL,
  ...
)
```

Arguments

`R1` — A `numeric` that provides the number of participants in the survey sample that were tested positive with both (medical) testing devices (and are, thus, members of the sub-population).

`R3` — A `numeric` that provides the number of participants in the survey sample that are tested positive only with the second testing device.

`n` — A `numeric` that provides the sample size.

`pi0` — A `numeric` that provides the prevalence or proportion of people (in the whole population) who are positive, as measured through a non-random, but systematic sampling (e.g. based on medical selection).

`gamma` — A `numeric` that is used to compute a (1 - gamma) confidence region for the proportion. Default value is `0.05`.

`alpha` — A `numeric` that provides the False Negative (FN) rate for the sample R. Default value is `0`.

`beta` — A `numeric` that provides the False Positive (FP) rate for the sample R. Default value is `0`.

`alpha0` — A `numeric` that corresponds to the probability that a random participant has been incorrectly declared positive through the nontransparent procedure. In most applications, this probability is likely very close to zero. Default value is `0`.

`V` A `numeric` that corresponds to the average of squared sampling weights. Default value is `NULL` and for the moment this method is currently only implemented for random sampling.

`...` Additional arguments.

Value

A `cpreval` object with the structure:

- estimate: Estimated proportion.
- sd: Estimated standard error of the estimator.
- ci_asym: Asymptotic confidence interval at the 1 - gamma confidence level.
- gamma: Confidence level (i.e. 1 - gamma) for confidence intervals.
- method: Estimation method (in this case marginal mle).
- measurement: A vector with (alpha0, alpha, beta).
- beta0: Estimated false negative rate of the official procedure.
- ci_beta0: Asymptotic confidence interval (1 - gamma confidence level) for beta0.
- boundary: A boolean variable indicating if the estimates falls at the boundary of the parameter space.
- pi0: Value of pi0 (input value).
- sampling: Type of sampling considered ("random" or "weighted").
- V: Average sum of squared sampling weights if weighted/stratified is used (otherwise NULL).
- n: Sample size.
- avar_beta0: Estimated asymptotic variance of beta0
- ...: Additional parameters

Author(s)

Stephane Guerrier, Maria-Pia Victoria-Feser, Christoph Kuzmics

Examples

```
# Samples without measurement error
X = sim_Rs(theta = 3/100, pi0 = 1/100, n = 1500, seed = 18)
conditional_mle(R1 = X$R1, R2 = X$R2, R3 = X$R3, R4 = X$R4, n = X$n, pi0 = X$pi0)

# With measurement error
X = sim_Rs(theta = 30/1000, pi0 = 10/1000, n = 1500, alpha0 = 0.001,
alpha = 0.01, beta0 = 0.05, beta = 0.05, seed = 18)
marginal_mle(R1 = X$R1, R3 = X$R3, n = X$n, pi0 = X$pi0)
marginal_mle(R1 = X$R1, R3 = X$R3, n = X$n, pi0 = X$pi0,
alpha0 = 0.001, alpha = 0.01, beta0 = 0.05, beta = 0.05)
```

`moment_estimator` *Compute moment-based estimator.*

Description

Proportion estimated using the moment-based estimator and confidence intervals based the asymptotic distribution of the estimator as well as the Clopper-Pearson approach.

Usage

```
moment_estimator(
  R3,
  n,
  pi0,
  gamma = 0.05,
  alpha = 0,
  beta = 0,
  alpha0 = 0,
  V = NULL,
  ...
)
```

Arguments

`R3` A `numeric` that provides the number of participants in the survey sample that are tested positive only with the second testing device.

`n` A `numeric` that provides the sample size.

`pi0` A `numeric` that provides the prevalence or proportion of people (in the whole population) who are positive, as measured through a non-random, but systematic sampling (e.g. based on medical selection).

`gamma` A `numeric` that is used to compute a (1 - gamma) confidence region for the proportion. Default value is `0.05`.

`alpha` A `numeric` that provides the False Negative (FN) rate for the sample R. Default value is `0`.

`beta` A `numeric` that provides the False Positive (FP) rate for the sample R. Default value is `0`.

`alpha0` A `numeric` that corresponds to the probability that a random participant has been incorrectly declared positive through the nontransparent procedure. In most applications, this probability is likely very close to zero. Default value is `0`.

`V` A `numeric` that corresponds to the average of squared sampling weights. Default value is `NULL`.

`...` Additional arguments.

Value

A `cpreval` object with the structure:

- estimate: Estimated proportion.
- sd: Estimated standard error of the estimator.
- ci_asym: Asymptotic confidence interval at the 1 - gamma confidence level.
- ci_cp: Confidence interval (1 - gamma confidence level) based on the Clopper-Pearson approach.
- gamma: Confidence level (i.e. 1 - gamma) for confidence intervals.
- method: Estimation method (in this case moment estimator).
- measurement: A vector with (alpha0, alpha, beta).
- beta0: Estimated false negative rate of the official procedure.
- ci_beta0: Asymptotic confidence interval (1 - gamma confidence level) for beta0.
- boundary: A boolean variable indicating if the estimates falls at the boundary of the parameter space.
- pi0: Value of pi0 (input value).
- sampling: Type of sampling considered ("random" or "weighted").
- V: Average sum of squared sampling weights if weighted/stratified is used (otherwise NULL).
- n: Sample size.
- avar_beta0: Estimated asymptotic variance of beta0
- ...: Additional parameters.

Author(s)

Stephane Guerrier, Maria-Pia Victoria-Feser, Christoph Kuzmics

Examples

```
# Samples without measurement error
X = sim_Rs(theta = 3/100, pi0 = 1/100, n = 1500, seed = 18)
moment_estimator(R3 = X$R3, n = X$n, pi0 = X$pi0)

# With measurement error
X = sim_Rs(theta = 3/100, pi0 = 1/100, n = 1500, alpha0 = 0.001,
alpha = 0.01, beta = 0.05, seed = 18)
moment_estimator(R3 = X$R3, n = X$n, pi0 = X$pi0)
moment_estimator(R3 = X$R3, n = X$n, pi0 = X$pi0, alpha0 = 0.001,
alpha = 0.01, beta = 0.05)
```

`sim_Rs` *Simulate data (R, R0, R1, R2, R3 and R4)*

Description

Simulation function for random variables of interest.

Usage

```
sim_Rs(theta, pi0, n, alpha0 = 0, alpha = 0, beta = 0, seed = NULL, ...)
```

Arguments

`theta` A `numeric` that provides the true prevalence of a given disease.

`pi0` A `numeric` that provides the prevalence or proportion of people (in the whole population) who are positive, as measured through a non-random, but systematic sampling (e.g. based on medical selection).

`n` A `numeric` that corresponds to the sample size.

`alpha0` A `numeric` that corresponds to the probability that a random participant has been incorrectly declared positive through the nontransparent procedure. In most applications, this probability is likely very close to zero. Default value is `0`.

`alpha` A `numeric` that provides the False Negative (FN) rate for the sample R. Default value is `0`.

`beta` A `numeric` that provides the False Positive (FP) rate for the sample R. Default value is `0`.

`seed` A `numeric` that provides the simulation seed. Default value is `NULL`.

`...` Additional arguments.

Value

A `cpreval_sim` object (`list`) with the structure:

- R: the number of participants in the survey sample that were tested positive.
- R0: the number of participants in the survey sample that were tested positive with the first testing device (and are, thus, members of the sub-population).
- R1: the number of participants in the survey sample that were tested positive with both (medical) testing devices (and are, thus, members of the sub-population).
- R2: the number of participants in the survey sample that are tested positive only with the first testing device (and are, thus, members of the sub-population).
- R3: the number of participants in the survey sample that are tested positive only with the second testing device.
- R4: the number of participants that are tested negative with the second testing device (and are either members of the sub-population and have tested negative with the first testing device or are not members of the sub-population).

- n: the sample size.
- alpha: the False Negative (FN) rate for the sample R.
- beta: the False Positive (FP) rate for the sample R.
- alpha0: the alpha0 probability (as defined above).
- ...: additional arguments.

Author(s)

Stephane Guerrier

Examples

```
# Samples without measurement error
sim_Rs(theta = 3/100, pi0 = 1/100, n = 1500, seed = 18)

# With measurement error
sim_Rs(theta = 3/100, pi0 = 1/100, n = 1500, alpha0 = 0,
alpha = 0.01, beta = 0.05, seed = 18)
```

`survey_mle` *Compute proportion in the survey sample (standard estimator)*

Description

Proportion estimated using the survey sample and confidence intervals based on the Clopper-Pearson and the standard asymptotic approach.

Usage

```
survey_mle(R, n, pi0 = 0, alpha = 0, beta = 0, gamma = 0.05, V = NULL, ...)
```

Arguments

`R` A `numeric` that provides the people of positive people in the sample.

`n` A `numeric` that provides the sample size.

`pi0` A `numeric` that provides the prevalence or proportion of people (in the whole population) who are positive, as measured through a non-random, but systematic sampling (e.g. based on medical selection). Default value is `0` and in this case this information is not used in the estimation procedure.

`alpha` A `numeric` that provides the False Negative (FN) rate for the sample R. Default value is `0`.

`beta` A `numeric` that provides the False Positive (FP) rate for the sample R. Default value is `0`.

`gamma` A `numeric` that is used to compute a (1 - gamma) confidence region for the proportion. Default value is `0.05`.

`V` A `numeric` that corresponds to the average of squared sampling weights. Default value is `NULL`.

`...` Additional arguments.

Value

A `cpreval` object with the structure:

- estimate: Estimated proportion.
- sd: Estimated standard error of the estimator.
- ci_asym: Asymptotic confidence interval at the 1 - gamma confidence level.
- gamma: Confidence level (i.e. 1 - gamma) for confidence intervals.
- method: Estimation method (in this case sample survey).
- measurement: A vector with (alpha0, alpha, beta).
- boundary: A boolean variable indicating if the estimates falls at the boundary of the parameter space.
- pi0: Value of pi0 (input value).
- sampling: Type of sampling considered ("random" or "weighted").
- V: Average sum of squared sampling weights if weighted/stratified is used (otherwise NULL).
- ...: Additional parameters.

Author(s)

Stephane Guerrier, Maria-Pia Victoria-Feser, Christoph Kuzmics

Examples

```
# Samples without measurement error
X = sim_Rs(theta = 30/1000, pi0 = 10/1000, n = 1500, seed = 18)
survey_mle(R = X$R, n = X$n)

# With measurement error
X = sim_Rs(theta = 30/1000, pi0 = 10/1000, n = 1500, alpha = 0.01, beta = 0.05, seed = 18)
survey_mle(R = X$R, n = X$n)
survey_mle(R = X$R, n = X$n, alpha = 0.01, beta = 0.05)
```

`update_prevalence` *Update prevalence using new case prevalence rates*

Description

Updated prevalence and confidence intervals using new case prevalence rates

Usage

```
update_prevalence(
  pi0_new,
  x,
  gamma = 0.05,
  print = NULL,
  plot = NULL,
  col_line = "#2e5dc1",
  col_ci = "#2E5DC133",
  ...
)
```

Arguments

pi0_new A numeric or vector of new case prevalence rates

x A cpreval object.

gamma A numeric that used to compute a (1 - gamma) confidence region for the proportion. Default value is 0.05.

print A boolean indicating whether or not the output should be print.

plot A boolean indicating whether or not a plot should be made.

col_line Color of the estimated prevalence.

col_ci Color of the estimated prevalence confidence interval.

... Additional arguments.

Value

A matrix object whose colunms corresponds to pi0, estimate, sd and CI.

Author(s)

Stephane Guerrier

Examples

```
# Austrian data (November 2020)
pi0 = 93914/7166167
data("covid19_austria")

# Weighted sampling
n = nrow(covid19_austria)
R1w = sum(covid19_austria$weights[covid19_austria$Y == 1 & covid19_austria$Z == 1])
R2w = sum(covid19_austria$weights[covid19_austria$Y == 0 & covid19_austria$Z == 1])
R3w = sum(covid19_austria$weights[covid19_austria$Y == 1 & covid19_austria$Z == 0])
R4w = sum(covid19_austria$weights[covid19_austria$Y == 0 & covid19_austria$Z == 0])

# Assumed measurement errors
alpha0 = 0
alpha = 1/100
```

```
beta = 10/100

# MME
mme = moment_estimator(R3 = R3w, n = n, pi0 = pi0, alpha = alpha, beta = beta,
                       alpha0 = alpha0, V = mean(covid19_austria$weights^2))

mme

# Update prevalence using a new pi0, say = 1.5%, instead of 1.31%
update_prevalence(1.5/100, mme)

pi0_new = seq(from = 0.005, to = 0.03, length.out = 100)
update_prevalence(pi0_new, mme)
```

Index