Package ‘epiR’

July 30, 2021

Version 2.0.33
Date 2021-07-30
Title Tools for the Analysis of Epidemiological Data
Author Mark Stevenson <mark.stevenson1@unimelb.edu.au> and Evan Sergeant <evansergeant@gmail.com> with contributions from Telmo Nunes, Cord Heuer, Jonathon Marshall, Javier Sanchez, Ron Thomson, Jeno Reiczigel, Jim Robison-Cox, Paola Sebastiani, Peter Solymos, Kazuki Yoshida, Geoff Jones, Sarah Pirikahu, Simon Firestone, Ryan Kyle, Johann Popp, Mathew Jay, Charles Reynard, Allison Cheung and Nagendra Singanallur.
Maintainer Mark Stevenson <mark.stevenson1@unimelb.edu.au>
Description Tools for the analysis of epidemiological and surveillance data. Contains functions for directly and indirectly adjusting measures of disease frequency, quantifying measures of association on the basis of single or multiple strata of count data presented in a contingency table, computation of confidence intervals around incidence risk and incidence rate estimates and sample size calculations for cross-sectional, case-control and cohort studies. Surveillance tools include functions to calculate an appropriate sample size for 1- and 2-stage representative freedom surveys, functions to estimate surveillance system sensitivity and functions to support scenario tree modelling analyses.
Depends R (>= 3.0.0), survival
Imports BiasedUrn, pander, methods, lubridate
Suggests MASS (>= 3.1-20), knitr, rmarkdown, RColorBrewer, ggplot2, plyr, rgdal, scales, spData, spatstat, foreign, maptools, rgeos, mapproj, tidyverse, kableExtra
VignetteBuilder knitr
License GPL (>= 2)
RoxygenNote 7.1.0
NeedsCompilation no
Repository CRAN
Date/Publication 2021-07-30 05:00:05 UTC
R topics documented:

epi.2by2 ................................................................. 4
epi.about ............................................................. 16
epi.asc ................................................................. 20
epi.betabuster ......................................................... 21
epi.blcm.paras ......................................................... 23
epi.bohning ........................................................... 25
epi.ccc ................................................................. 26
epi.conf ............................................................... 31
epi.convgrid .......................................................... 36
epi.cp ................................................................. 37
epi.cpresids .......................................................... 38
epi.descriptives ...................................................... 40
epi.dgamma ............................................................ 41
epi.directadj .......................................................... 42
epi.dms ................................................................. 46
epi.dsl ................................................................. 47
epi.edr ................................................................. 49
epi.empbayes ......................................................... 51
epi.epidural ......................................................... 52
epi.herdtest ........................................................... 53
epi.incin .............................................................. 54
epi.indirectadj ....................................................... 55
epi.insthaz ............................................................ 58
epi.interaction ....................................................... 60
epi.iv ................................................................. 63
epi.kappa .............................................................. 66
epi.ltd ................................................................. 69
epi.mh ................................................................. 71
epi.nomogram ........................................................ 73
epi.occc ............................................................... 74
epi.offset ............................................................. 76
epi.pooled ............................................................ 77
epi.popsize ........................................................... 78
epi.prcc .............................................................. 79
epi.prev .............................................................. 81
epi.psi ............................................................... 83
epi.RtoBUGS .......................................................... 85
epi.SClip ............................................................. 86
epi.smd ............................................................... 87
epi.smr ............................................................... 89
epi.sscc .............................................................. 91
epi.ssclus1estb ...................................................... 95
epi.ssclus1estc ...................................................... 97
epi.ssclus2estb ...................................................... 99
epi.ssclus2estc ..................................................... 101
epi.sscohortc ....................................................... 102
R topics documented:

epi.sscohortt ........................................ 106
epi.sscompb ........................................ 108
epi.sscompc ........................................ 110
epi.sscomps ........................................ 113
epi.ssdetect ........................................ 115
epi.ssdxtest .......................................... 117
epi.ssequb .......................................... 119
epi.ssequc .......................................... 121
epi.ssninfb .......................................... 124
epi.ssninfc .......................................... 126
epi.sssimpleestb .................................... 129
epi.sssimpleestc ................................... 131
epi.issstrataestb .................................. 133
epi.issstrataestc .................................. 134
epi.sssupb .......................................... 136
epi.sssupc .......................................... 138
epi.ssssectn ........................................ 140
epi.tests ........................................... 142
rsu.adjrisk .......................................... 146
rsu.dxtest ........................................... 148
rsu.epinf ........................................... 151
rsu.pfree.equ ....................................... 152
rsu.pfree.rs ......................................... 154
rsu.pstar ............................................ 158
rsu.sep ............................................... 160
rsu.sep.cens ........................................ 161
rsu.sep.pass ......................................... 162
rsu.sep.rb .......................................... 163
rsu.sep.rb1rf ....................................... 165
rsu.sep.rb2rf ....................................... 167
rsu.sep.rb2st ....................................... 169
rsu.sep.rbvarse .................................... 172
rsu.sep.rs ........................................... 173
rsu.sep.rs2st ....................................... 174
rsu.sep.rsfreecalc ................................ 176
rsu.sep.rsmult ..................................... 177
rsu.sep.rspool ...................................... 179
rsu.sep.rsvarse .................................... 180
rsu.spp.rs ........................................... 182
rsu.sspfree.rs ..................................... 183
rsu.sssep.rb2st1rf ................................ 185
rsu.sssep.rb2st2rf ................................ 186
rsu.sssep.rbmrg .................................... 188
rsu.sssep.rbsrg .................................... 190
rsu.sssep.rs ......................................... 191
rsu.sssep.rs2st ..................................... 193
rsu.sssep.rsfreecalc ................................ 195
rsu.sssep.rspool .................................... 197
Description

Computes summary measures of risk and a chi-squared test for difference in the observed proportions from count data presented in a 2 by 2 table. With multiple strata the function returns crude and Mantel-Haenszel adjusted measures of association and chi-squared tests of homogeneity.

Usage

epi.2by2(dat, method = "cohort.count", conf.level = 0.95, units = 100,
   interpret = FALSE, outcome = "as.columns")

Arguments

dat a vector of length four, an object of class table or an object of class grouped_df from package dplyr containing the individual cell frequencies (see below).

method a character string indicating the study design on which the tabular data has been based. Options are cohort.count, cohort.time, case.control, or cross.sectional. Based on the study design specified by the user, appropriate measures of association, measures of effect in the exposed and measures of effect in the population are returned by the function.

conf.level magnitude of the returned confidence intervals. Must be a single number between 0 and 1.

units multiplier for prevalence and incidence (risk or rate) estimates.

interpret logical. If TRUE interpretive statements are appended to the printepi.2by2 object.

outcome a character string indicating how the outcome variable is represented in the contingency table. Options are as.columns (outcome as columns) or as.rows (outcome as rows).

x, object an object of class epi.2by2.

... Ignored.

Details

Where method is cohort.count, case.control, or cross.sectional and outcome = as.columns the required 2 by 2 table format is:
Where method is cohort.time and outcome = as.columns the required 2 by 2 table format is:

<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expose +</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Expose -</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

A summary of the methods used for each of the confidence interval calculations in this function is as follows:

**Value**

An object of class epi.2by2 comprised of:

- **method**: character string returning the study design specified by the user.
- **n.strata**: number of strata.
- **conf.level**: magnitude of the returned confidence intervals.
- **interp**: logical. Are interpretative statements included?
- **units**: character string listing the outcome measure units.
- **tab**: a data frame comprised of of the contingency table data.
- **massoc.summary**: a data frame listing the computed measures of association, measures of effect in the exposed and measures of effect in the population and their confidence intervals.
- **massoc.interp**: a data frame listing the interpretive statements for each computed measure of association.
- **massoc.detail**: a list comprised of the computed measures of association, measures of effect in the exposed and measures of effect in the population. See below for details.

When method equals cohort.count the following measures of association, measures of effect in the exposed and measures of effect in the population are returned:
RR Wald, Taylor and score confidence intervals for the incidence risk ratios for each strata. Wald, Taylor and score confidence intervals for the crude incidence risk ratio. Wald confidence interval for the Mantel-Haenszel adjusted incidence risk ratio.

OR Wald, score, Cornfield and maximum likelihood confidence intervals for the odds ratios for each strata. Wald, score, Cornfield and maximum likelihood confidence intervals for the crude odds ratio. Wald confidence interval for the Mantel-Haenszel adjusted odds ratio.

ARisk Wald and score confidence intervals for the attributable risk (risk difference) for each strata. Wald, Sato and Greenland-Robins confidence intervals for the Mantel-Haenszel adjusted attributable risk.

PARisk Wald and Pirikahu confidence intervals for the population attributable risk for each strata. Wald and Pirikahu confidence intervals for the crude population attributable risk. The Pirikahu confidence intervals are calculated using the delta method.

AFRisk Wald confidence intervals for the attributable fraction for each strata. Wald confidence intervals for the crude attributable fraction.

PAFRisk Wald confidence intervals for the population attributable fraction for each strata. Wald confidence interval for the crude population attributable fraction.

chisq.strata chi-squared test for difference in exposed and non-exposed proportions for each strata.

chisq.crude chi-squared test for difference in exposed and non-exposed proportions across all strata.

chisq.mh Mantel-Haenszel chi-squared test that the combined odds ratio estimate is equal to 1.

RR.homog Mantel-Haenszel (Woolf) test of homogeneity of the individual strata incidence risk ratios.

OR.homog Mantel-Haenszel (Woolf) test of homogeneity of the individual strata odds ratios.

When method equals cohort.time the following measures of association and effect are returned:

IRR Wald confidence interval for the incidence rate ratios for each strata. Wald confidence interval for the crude incidence rate ratio. Wald confidence interval for the Mantel-Haenszel adjusted incidence rate ratio.

ARate Wald confidence interval for the attributable rate for each strata. Wald confidence interval for the crude attributable rate. Wald confidence interval for the Mantel-Haenszel adjusted attributable rate.

PARate Wald confidence interval for the population attributable rate for each strata. Wald confidence interval for the crude population attributable rate.

AFRate Wald confidence interval for the attributable fraction for each strata. Wald confidence interval for the crude attributable fraction.

PAFRate Wald confidence interval for the population attributable fraction for each strata. Wald confidence interval for the crude population attributable fraction.
chisq.strata  chi-squared test for difference in exposed and non-exposed proportions for each strata.
chisq.crude  chi-squared test for difference in exposed and non-exposed proportions across all strata.
chisq.mh    Mantel-Haenszel chi-squared test that the combined odds ratio estimate is equal to 1.

When method equals case.control the following measures of association and effect are returned:

OR       Wald, score, Cornfield and maximum likelihood confidence intervals for the odds ratios for each strata. Wald, score, Cornfield and maximum likelihood confidence intervals for the crude odds ratio. Wald confidence interval for the Mantel-Haenszel adjusted odds ratio.
ARisk    Wald and score confidence intervals for the attributable risk for each strata. Wald and score confidence intervals for the crude attributable risk. Wald, Sato and Greenland-Robins confidence intervals for the Mantel-Haenszel adjusted attributable risk.
PARisk   Wald and Pirikahu confidence intervals for the population attributable risk for each strata. Wald and Pirikahu confidence intervals for the crude population attributable risk.
AFest    Wald confidence intervals for the estimated attributable fraction for each strata. Wald confidence intervals for the crude estimated attributable fraction.
PAFest   Wald confidence intervals for the population estimated attributable fraction for each strata. Wald confidence intervals for the crude population estimated attributable fraction.
chisq.strata  chi-squared test for difference in exposed and non-exposed proportions for each strata.
chisq.crude  chi-squared test for difference in exposed and non-exposed proportions across all strata.
chisq.mh    Mantel-Haenszel chi-squared test that the combined odds ratio estimate is equal to 1.
OR.homog   Mantel-Haenszel (Woolf) test of homogeneity of the individual strata odds ratios.

When method equals cross-sectional the following measures of association and effect are returned:

PR       Wald, Taylor and score confidence intervals for the prevalence ratios for each strata. Wald, Taylor and score confidence intervals for the crude prevalence ratio. Wald confidence interval for the Mantel-Haenszel adjusted prevalence ratio.
OR       Wald, score, Cornfield and maximum likelihood confidence intervals for the odds ratios for each strata. Wald, score, Cornfield and maximum likelihood confidence intervals for the crude odds ratio. Wald confidence interval for the Mantel-Haenszel adjusted odds ratio.
ARisk
Wald and score confidence intervals for the attributable risk for each strata.
Wald and score confidence intervals for the crude attributable risk. Wald, Sato and
Greenland-Robins confidence intervals for the Mantel-Haenszel adjusted
attributable risk.

PARisk
Wald and Pirikahu confidence intervals for the population attributable risk for
each strata. Wald and Pirikahu confidence intervals for the crude population
attributable risk.

AFRisk
Wald confidence intervals for the attributable fraction for each strata. Wald con-
fidence intervals for the crude attributable fraction.

PAFRisk
Wald confidence intervals for the population attributable fraction for each strata.
Wald confidence intervals for the crude population attributable fraction.

chisq.strata
chi-squared test for difference in exposed and non-exposed proportions for each
strata.

chisq.crude
chi-squared test for difference in exposed and non-exposed proportions across
all strata.

chisq.mh
Mantel-Haenszel chi-squared test that the combined odds ratio estimate is equal
to 1.

PR.homog
Mantel-Haenszel (Woolf) test of homogeneity of the individual strata prevalence
ratios.

OR.homog
Mantel-Haenszel (Woolf) test of homogeneity of the individual strata odds ra-
tios.

The point estimates of the wald, score and cfield odds ratios are calculated using the cross product
method. Method mle computes the conditional maximum likelihood estimate of the odds ratio.

Confidence intervals for the Cornfield (cfield) odds ratios are computed using the hypergeometric
distribution and computation times are extremely slow when the cell frequencies are large. For this
reason, Cornfield confidence intervals are only calculated if the total number of event frequencies
is less than 500. Maximum likelihood estimates of the odds ratio are only calculated when the total
number of observations is less than 2E09.

If the Haldane-Anscombe (Haldane 1940, Anscombe 1956) correction is applied (i.e. addition of
0.5 to each cell of the 2 by 2 table when at least one of the cell frequencies is zero) Cornfield
(cfield) odds ratios are not computed.

The Mantel-Haenszel chi-squared test that the combined odds ratio estimate is equal to 1 uses a
two-sided test without continuity correction.

Interpretive statements for the number needed to treat to benefit (NNTB) and number needed to treat
to harm (NNTH) follow the approach described by Altman (1998). See the examples for details.

**Note**

Measures of association include the prevalence ratio, the incidence risk ratio, the incidence rate
ratio and the odds ratio. The incidence risk ratio is the ratio of the incidence risk of disease in
the exposed group to the incidence risk of disease in the unexposed group. The odds ratio (also
known as the cross-product ratio) is an estimate of the incidence risk ratio. When the incidence of
an outcome in the study population is low (say, less than 5%) the odds ratio will provide a reliable
estimate of the incidence risk ratio. The more frequent the outcome becomes, the more the odds
ratio will overestimate the incidence risk ratio when it is greater than than 1 or underestimate the incidence risk ratio when it is less than 1.

Measures of effect in the exposed include the attributable risk (or prevalence) and the attributable fraction. The attributable risk is the risk of disease in the exposed group minus the risk of disease in the unexposed group. The attributable risk provides a measure of the absolute increase or decrease in risk associated with exposure. The attributable fraction is the proportion of study outcomes in the exposed group that is attributable to exposure.

Measures of effect in the population include the population attributable risk (or prevalence) and the population attributable fraction (also known as the aetiologic fraction). The population attributable risk is the risk of the study outcome in the population that may be attributed to exposure. The population attributable fraction is the proportion of the study outcomes in the population that is attributable to exposure.

Point estimates and confidence intervals for the prevalence ratio and incidence risk ratio are calculated using Wald (Wald 1943) and score methods (Miettinen and Nurminen 1985). Point estimates and confidence intervals for the incidence rate ratio are calculated using the exact method described by Kirkwood and Sterne (2003) and Juul (2004). Point estimates and confidence intervals for the population attributable risk are calculated using formulae provided by Rothman and Greenland (1998, p 271) and Pirikahu (2014). Point estimates and confidence intervals for the population attributable fraction are calculated using formulae provided by Jewell (2004, p 84 - 85). Point estimates and confidence intervals for the Mantel-Haenszel adjusted attributable risk are calculated using formulae provided by Klingenberg (2014).

Wald confidence intervals are provided in the summary table simply because they are widely used and would be familiar to most users.

The Mantel-Haenszel adjusted measures of association are valid when the measures of association across the different strata are similar (homogenous), that is when the test of homogeneity of the odds (risk) ratios is not significant.

The Mantel-Haenszel (Woolf) test of homogeneity of the odds ratio are based on Jewell (2004, p 152 - 158). Thanks to Jim Robison-Cox for sharing his implementation of these functions.

Author(s)

Mark Stevenson (Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Australia), Cord Heuer (EpiCentre, IVABS, Massey University, Palmerston North, New Zealand), Jim Robison-Cox (Department of Math Sciences, Montana State University, Montana, USA), Kazuki Yoshida (Brigham and Women’s Hospital, Boston Massachusetts, USA) and Simon Firestone (Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Australia). Thanks to Ian Dohoo for numerous helpful suggestions to improve the documentation for this function.

References


Haldane J (1940). The mean and variance of the moments of chi square, when used as a test of homogeneity, when expectations are small. Biometrika 29, 133 - 143.


Wald A (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society 54: 426 - 482.

Examples

## EXAMPLE 1:
## A cross sectional study investigating the relationship between dry cat food (DCF) and feline urologic syndrome (FUS) was conducted (Willeberg 1977). Counts of individuals in each group were as follows:

## DCF-exposed cats (cases, non-cases) 13, 2163
## Non DCF-exposed cats (cases, non-cases) 5, 3349

## Outcome variable (FUS) as columns:
dat.v01 <- c(13,2163,5,3349); dat.v01
epi.2by2(dat = dat.v01, method = "cross.sectional", conf.level = 0.95, units = 100, interpret = FALSE, outcome = "as.columns")

## Outcome variable (FUS) as rows:
dat.v01 <- c(13,5,2163,3349); dat.v01
epi.2by2(dat = dat.v01, method = "cross.sectional", conf.level = 0.95, units = 100, interpret = FALSE, outcome = "as.rows")

## The prevalence of FUS in DCF exposed cats was 4.01 (95% CI 1.43 to 11.23) times greater than the prevalence of FUS in non-DCF exposed cats.

## In DCF exposed cats, 75% (95% CI 30% to 91%) of the FUS cases were attributable to DCF.

## Fifty-four percent of FUS cases in the population was attributable to DCF (95% CI 4% to 78%).

## EXAMPLE 2:
## This example shows how the table function in base R can be used to pass data to epi.2by2. Here we use the birthwt data set from the MASS package.

library(MASS)
dat.df02 <- birthwt; head(dat.df02)

## Generate a table of cell frequencies. First, set the outcome and exposure
## as factors and set their levels appropriately so the frequencies in the
## 2 by 2 table come out in the conventional format:
dat.df02$low <- factor(dat.df02$low, levels = c(1,0))
dat.df02$smoke <- factor(dat.df02$smoke, levels = c(1,0))
dat.df02$race <- factor(dat.df02$race, levels = c(1,2,3))
dat.tab02 <- table(dat.df02$smoke, dat.df02$low, dnn = c("Smoke", "Low BW"))
print(dat.tab02)

## Compute the odds ratio and other measures of association:
epi.2by2(dat = dat.tab02, method = "cohort.count", conf.level = 0.95,
units = 100, interpret = FALSE, outcome = "as.columns")

## The odds of having a low birth weight child for smokers was 2.02
## (95% CI 1.08 to 3.78) times greater than the odds of having a low birth
## weight child for non-smokers.

## Stratify by race:
dat.tab02 <- table(dat.df02$smoke, dat.df02$low, dat.df02$race,
dnn = c("Smoke", "Low BW", "Race"))
print(dat.tab02)

## Compute the crude odds ratio, the Mantel-Haenszel adjusted odds ratio
## and other measures of association:
dat.epi02 <- epi.2by2(dat = dat.tab02, method = "cohort.count",
conf.level = 0.95, units = 100, interpret = FALSE, outcome = "as.columns")
print(dat.epi02)

## The Mantel-Haenszel test of homogeneity of the strata odds ratios is not
## significant (chi square test statistic 2.800; df 2; p-value = 0.25).
## We accept the null hypothesis and conclude that the odds ratios for
## each strata of race are the same.

## After accounting for the confounding effect of race, the odds of
## having a low birth weight child for smokers was 3.09 (95% CI 1.49 to 6.39)
## times that of non-smokers.

## Compare the Greenland-Robins confidence intervals for the Mantel-Haenszel
## adjusted attributable risk with the Wald confidence intervals for the
## Mantel-Haenszel adjusted attributable risk:
dat.epi02$massoc.detail$ARisk.mh.green
dat.epi02$massoc.detail$ARisk.mh.wald

## How many mothers need to stop smoking to avoid one low birth weight baby?
dat.epi02$massoc.interp$text[dat.epi02$massoc.interp$var ==
"NNTB NNTH (crude)""]

## If we don’t account for confounding the number of mothers that need to
## stop smoking to avoid one low birth weight baby (NNTB) is
## 7 (95% CI 3 to 62).
After accounting for the confounding effect of race the number of mothers that need to stop smoking to avoid one low birth weight baby (NNTB) is 5 (95% CI 2 to 71).

Now turn dat.tab02 into a data frame where the frequencies of individuals in each exposure-outcome category are provided. Often your data will be presented in this summary format:

```r
dat.df02 <- data.frame(dat.tab02); head(dat.df02)
```

Re-format dat.df02 (a summary count data frame) into tabular format using the xtabs function:

```r
dat.tab02 <- xtabs(Freq ~ Smoke + Low.BW + Race, data = dat.df02)
print(dat.tab02)
```

# dat02.tab can now be passed to epi.2by2:

```r
dat.epi02 <- epi.2by2(dat = dat.tab02, method = "cohort.count", conf.level = 0.95, units = 100, interpret = FALSE, outcome = "as.columns")
print(dat.epi02)
```

The Mantel-Haenszel adjusted odds ratio is 3.09 (95% CI 1.49 to 6.39). The ratio of the crude odds ratio to the Mantel-Haenszel adjusted odds ratio is 0.66.

What are the Cornfield confidence limits, the maximum likelihood confidence limits and the score confidence limits for the crude odds ratio?

```r
dat.epi02$massoc.detail$OR.crude.cfield
dat.epi02$massoc.detail$OR.crude.mle
dat.epi02$massoc.detail$OR.crude.score
```

Cornfield: 2.02 (95% CI 1.07 to 3.79)
Maximum likelihood: 2.01 (1.03 to 3.96)
Score: 2.02 (95% CI 1.08 to 3.77)

Plot the individual strata-level odds ratios and compare them with the Mantel-Haenszel adjusted odds ratio.

```r
library(ggplot2); library(scales)
nstrata <- 1:dim(dat.tab02)[3]
strata.lab <- paste("Strata ", nstrata, sep = "")
y.at <- c(nstrata, max(nstrata) + 1)
y.lab <- c("M-H", strata.lab)
x.at <- c(0.25,0.5,1,2,4,8,16,32)
or.p <- c(dat.epi02$massoc.detail$OR.mh$est, dat.epi02$massoc.detail$OR.strata.cfield$est)
or.l <- c(dat.epi02$massoc.detail$OR.mh$lower, dat.epi02$massoc.detail$OR.strata.cfield$lower)
or.u <- c(dat.epi02$massoc.detail$OR.mh$upper, dat.epi02$massoc.detail$OR.strata.cfield$upper)
```
dat.df02 <- data.frame(y.at, y.lab, or.p, or.l, or.u)

ggplot(data = dat.df02, aes(x = or.p, y = y.at)) +
  geom_point() +
  geom_errorbarh(aes(xmax = or.l, xmin = or.u, height = 0.2)) +
  labs(x = "Odds ratio", y = "Strata") +
  scale_x_continuous(trans = log2_trans(), breaks = x.at,
    limits = c(0.25, 32)) +
  scale_y_continuous(breaks = y.at, labels = y.lab) +
  geom_vline(xintercept = 1, lwd = 1) +
  coord_fixed(ratio = 0.75 / 1) +
  theme(axis.title.y = element_text(vjust = 0))

## End(Not run)

## EXAMPLE 3:
## Same as Example 2 but showing how a 2 by 2 contingency table can be prepared
## using tidyverse:

## Not run:
library(MASS); library(tidyverse)

dat.df03 <- birthwt; head(dat.df03)

dat.tab03 <- dat.df03 %>%
  mutate(low = factor(low, levels = c(1, 0), labels = c("yes", "no"))) %>%
  mutate(smoke = factor(smoke, levels = c(1, 0), labels = c("yes", "no"))) %>%
  mutate(race = factor(race)) %>%
  group_by(race, smoke, low) %>%
  summarise(n = n())

dat.tab03

## View the data in conventional 2 by 2 table format:
pivot_wider(dat.tab03, id_cols = c(race, smoke),
  names_from = low, values_from = n)

dat.epi03 <- epi.2by2(dat = dat.tab03, method = "cohort.count",
  conf.level = 0.95, units = 100, interpret = FALSE, outcome = "as.columns")

dat.epi03

## End(Not run)

## The Mantel-Haenszel test of homogeneity of the strata odds ratios is not
## significant (chi square test statistic 2.800; df 2; p-value = 0.25).
## We accept the null hypothesis and conclude that the odds ratios for
## each strata of race are the same.

## After accounting for the confounding effect of race, the odds of
## having a low birth weight child for smokers was 3.09 (95% CI 1.49 to 6.39)
## times that of non-smokers.
## EXAMPLE 4:
A study was conducted by Feychting et al (1998) comparing cancer occurrence among the blind with occurrence among those who were not blind but had severe visual impairment. From these data we calculate a cancer rate of 136/22050 person-years among the blind compared with 1709/127650 person-years among those who were visually impaired but not blind.

```r
dat.v04 <- c(136,22050,1709,127650)
dat.epi04 <- epi.2by2(dat = dat.v04, method = "cohort.time", conf.level = 0.95, units = 1000, interpret = FALSE, outcome = "as.columns")
summary(dat.epi04)$massoc.detail$ARate.strata.wald
```

The incidence rate of cancer was 7.22 (95% CI 6.00 to 8.43) cases per 1000 person-years less in the blind, compared with those who were not blind but had severe visual impairment.

```r
round(summary(dat.epi04)$massoc.detail$IRR.strata.wald, digits = 2)
```

The incidence rate of cancer in the blind group was less than half that of the comparison group (incidence rate ratio 0.46, 95% CI 0.38 to 0.55).

## EXAMPLE 5:
A study has been conducted to assess the effect of a new treatment for mastitis in dairy cows. Eight herds took part in the study. The following data were obtained. The vectors ai, bi, ci and di list (for each herd) the number of cows in the E+D+, E+D-, E-D+ and E-D- groups, respectively.

```r
hid <- 1:8
ai <- c(23,10,20,5,14,6,10,3)
bi <- c(10,2,1,2,2,2,3,0)
ci <- c(3,2,3,2,1,3,3,2)
di <- c(6,4,3,2,6,3,1,1)
dat.df05 <- data.frame(hid, ai, bi, ci, di)
head(dat.df05)
```

Re-format data into a format suitable for epi.2by2:

```r
hid <- rep(1:8, times = 4)
exp <- factor(rep(c(1,0,1,0), each = 8), levels = c(1,0))
out <- factor(rep(c(1,0,1,0), each = 8), levels = c(1,0))
dat.df05 <- data.frame(hid, exp, out, n = c(ai,bi,ci,di))
dat.tab05 <- xtabs(n ~ exp + out + hid, data = dat.df05)
print(dat.tab05)
```

epi.2by2(dat = dat.tab05, method = "cohort.count", conf.level = 0.95, units = 1000, interpret = FALSE, outcome = "as.columns")

The Mantel-Haenszel test of homogeneity of the strata odds ratios is not significant (chi square test statistic 5.276; df 7; p-value = 0.63).

We accept the null hypothesis and conclude that the odds ratios for each strata of herd are the same.
## After adjusting for the effect of herd, compared to untreated cows, treatment
## increased the odds of recovery by a factor of 5.97 (95% CI 2.72 to 13.13).
## End(Not run)

### Description

Tools for the analysis of epidemiological data.

### Usage

epi.about()

### Details


### FUNCTIONS AND DATASETS

The following is a summary of the main functions and datasets in the epiR package. An alphabetical list of all functions and datasets is available by typing `library(help = epiR)`.

For further information on any of these functions, type `help(name)` or `?name` where name is the name of the function or dataset.

For details on how to use epiR for routine epidemiological work start R, type `help.start()` to open the help browser and navigate to Packages > epiR > Vignettes.

### CONTENTS:

The functions in epiR can be categorised into two main groups: tools for epidemiological analysis and tools for the analysis of surveillance data. A summary of the package functions is as follows:

#### I. EPIDEMIOLOGY

1. Descriptive statistics:
   - epi.conf
     - Confidence intervals.
   - epi.descriptives
     - Descriptive statistics.

2. Measures of health and measures of association:
   - epi.directadj
     - Directly adjusted incidence rate estimates.
epi.about

epi.edr Compute estimated dissemination ratios from outbreak event data.
epi.empbayes Empirical Bayes estimates of observed event counts.
epi.indirectadj Indirectly adjusted incidence risk estimates.
epi.insthaz Instantaneous hazard estimates based on Kaplan-Meier survival estimates.
epi.2by2 Measures of association from data presented in a 2 by 2 table.

3. Diagnostic tests:

epi.betabuster An R version of Wes Johnson and Chun-Lung Su’s Betabuster.
epi.herdtest Estimate the characteristics of diagnostic tests applied at the herd (group) level.
epi.nomogram Compute the post-test probability of disease given characteristics of a diagnostic test.
epi.pooled Estimate herd test characteristics when samples are pooled.
epi.prev Compute the true prevalence of a disease in a population on the basis of an imperfect test.
epi.tests Sensitivity, specificity and predictive value of a diagnostic test.

4. Meta-analysis:

epi.dsl Mixed-effects meta-analysis of binary outcome data using the DerSimonian and Laird method.
epi.iv Fixed-effects meta-analysis of binary outcome data using the inverse variance method.
epi.mh Fixed-effects meta-analysis of binary outcome data using the Mantel-Haenszel method.
epi.smd Fixed-effects meta-analysis of continuous outcome data using the standardised mean difference method.

5. Regression analysis tools:

epi.cp Extract unique covariate patterns from a data set.
epi.cpresids Compute covariate pattern residuals from a logistic regression model.
epi.interaction Relative excess risk due to interaction in a case-control study.

6. Data manipulation tools:

epi.asc Write matrix to an ASCII raster file.
epi.convgrid Convert British National Grid georeferences to easting and northing coordinates.
epi.dms Convert decimal degrees to degrees, minutes and seconds and vice versa.
epi.ltd Calculate lactation to date and standard lactation (that is, 305 or 270 day) milk yields.
epi.offset Create an offset vector based on a list suitable for WinBUGS.
epi.RtoBUGS Write data from an R list to a text file in WinBUGS-compatible format.

7. Sample size calculations:

The naming convention for the sample size functions in epiR is: epi.ss (sample size) + an abbreviation to represent the sampling design (e.g. simple, strata, clus1, clus2) + an abbreviation of the objectives of the study (est when you want to estimate a population parameter or comp when you want to compare two groups) + a single letter defining the outcome variable type (b for binary, c for continuous and s for survival data).

epi.sssimpleestb Sample size to estimate a binary outcome using simple random sampling.
epi.sssimpleestc Sample size to estimate a continuous outcome using simple random sampling.
epi.ssstrataestb Sample size to estimate a binary outcome using stratified random sampling.
epi.ssstrataestc Sample size to estimate a continuous outcome using stratified random sampling.
epi.ssclus1estb Sample size to estimate a binary outcome using one-stage cluster sampling.
epi.ssclus1estc Sample size to estimate a continuous outcome using one-stage cluster sampling.
epi.ssclus2estb Sample size to estimate a binary outcome using two-stage cluster sampling.
epi.ssclus2estc Sample size to estimate a continuous outcome using two-stage cluster sampling.
epi.sxxsectn Sample size, power or detectable prevalence ratio for a cross-sectional study.
epi.sscohortc Sample size, power or detectable risk ratio for a cohort study using count data.
epi.sscohorttt Sample size, power or detectable risk ratio for a cohort study using time at risk data.
epi.sscc Sample size, power or detectable odds ratio for case-control studies.
epi.ssscompb Sample size, power and detectable risk ratio when comparing binary outcomes.
epi.ssscompc Sample size, power and detectable risk ratio when comparing continuous outcomes.
epi.ssscomps Sample size, power and detectable hazard when comparing time to event.
epi.ssequb Sample size for a parallel equivalence trial, binary outcome.
epi.ssequc Sample size for a parallel equivalence trial, continuous outcome.
epi.sssupb Sample size for a parallel superiority trial, binary outcome.
epi.sssupc Sample size for a parallel superiority trial, continuous outcome.
epi.ssninfb Sample size for a non-inferiority trial, binary outcome.
epi.ssninfc Sample size for a non-inferiority trial, continuous outcome.
epi.ssdetect Sample size to detect an event.
epi.ssdxtest Sample size to validate a diagnostic test in the absence of a gold standard.

8. Miscellaneous functions:
   epi.prcc Compute partial rank correlation coefficients.
   epi.psi Compute proportional similarity indices.

9. Data sets:
   epi.epidural Rates of use of epidural anaesthesia in trials of caregiver support.
II. SURVEILLANCE

Below, SSe stands for surveillance system sensitivity. That is, the average probability that a surveillance system (as a whole) will return a positive surveillance outcome, given disease is present in the population at a level equal to or greater than a specified design prevalence.

1. Representative sampling — sample size:

   - rsu.sspfree.rs: Defined probability of disease freedom.
   - rsu.sssep.rs: SSe, perfect test specificity.
   - rsu.sssep.rs2st: SSe, two stage sampling.
   - rsu.sssep.rsfreecalc: SSe, imperfect test specificity.
   - rsu.sssep.rspool: SSe, pooled sampling.

2. Representative sampling — surveillance system sensitivity and specificity:

   - rsu.sep.rs: SSe, representative sampling.
   - rsu.sep.rs2st: SSe, representative two-stage sampling.
   - rsu.sep.rsmult: SSe, representative multiple surveillance components.
   - rsu.sep.rsfreecalc: SSe, imperfect test specificity.
   - rsu.sep.rspool: SSe, representative pooled sampling.
   - rsu.sep.rsvarse: SSe, varying surveillance unit sensitivity.
   - rsu.spp.rs: Surveillance system specificity.

3. Representative sampling — probability of disease freedom:

   - rsu.pfree.rs: Probability of disease freedom for a single or multiple time periods.
   - rsu.pfree.equ: Equilibrium probability of disease freedom.

4. Risk-based sampling — sample size:

   - rsu.sssep.rbsrg: SSe, single sensitivity for each risk group.
   - rsu.sssep.rbmrg: SSe, multiple sensitivities within risk groups.
   - rsu.sssep.rb2st1rf: SSe, 2 stage sampling, 1 risk factor.
   - rsu.sssep.rb2st2rf: SSe, 2 stage sampling, 2 risk factors.

5. Risk-based sampling — surveillance system sensitivity and specificity:

   - rsu.sep.rb: SSe, risk-based sampling.
   - rsu.sep.rb1rf: SSe, risk-based sampling, 1 risk factor.
   - rsu.sep.rb2rf: SSe, risk-based sampling, 2 risk factors.
   - rsu.sep.rbvarse: SSe, risk-based sampling, varying unit sensitivity.
   - rsu.sep.rb2st: SSe, 2-stage risk-based sampling.

6. Risk-based sampling — probability of disease freedom:
rsu.pfree.equ  Equilibrium probability of disease freedom.

7. Census sampling — surveillance system sensitivity:

    rsu.sep.cens  SSe, census sampling.

8. Passive surveillance — surveillance system sensitivity:

    rsu.sep.pass  SSe, passive surveillance.

9. Miscellaneous functions:

    rsu.adjrisk  Adjusted risk values.
    rsu.dxtest  Series and parallel diagnostic test interpretation.
    rsu.epinf  Effective probability of disease.
    rsu.pstar  Design prevalence back calculation.
    rsu.sep  Probability disease is less than specified design prevalence.

Author(s)

Mark Stevenson (<mark.stevenson1@unimelb.edu.au>), Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville Victoria 3010, Australia.

Evan Sergeant (<evansergeant@gmail.com>), Ausvet Pty Ltd, Level 1 34 Thynne St, Bruce ACT 2617, Australia.

Simon Firestone, Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville Victoria 3010, Australia.

Telmo Nunes, UISEE/DETSA, Faculdade de Medicina Veterinaria — UTL, Rua Prof. Cid dos Santos, 1300 - 477 Lisboa Portugal.

Javier Sanchez, Atlantic Veterinary College, University of Prince Edward Island, Charlottetown Prince Edward Island, C1A 4P3, Canada.

Ron Thornton, Ministry for Primary Industries New Zealand, PO Box 2526 Wellington, New Zealand.

epi.asc  Write matrix to an ASCII raster file

Description

Writes a data frame to an ASCII raster file, suitable for display in a Geographic Information System.

Usage

epi.asc(dat, file, xllcorner, yllcorner, cellsize, na = -9999)
Arguments

dat a matrix with data suitable for plotting using the `image` function.
file character string specifying the name and path of the ASCII raster output file.
xllcorner the easting coordinate corresponding to the lower left hand corner of the matrix.
yllcorner the northing coordinate corresponding to the lower left hand corner of the matrix.
cellsize number, defining the size of each matrix cell.
na scalar, defines null values in the matrix. NAs are converted to this value.

Value

Writes an ASCII raster file (typically with *.asc extension), suitable for display in a Geographic Information System.

Note

The `image` function in R rotates tabular data counter clockwise by 90 degrees for display. A matrix of the form:

\[
\begin{pmatrix}
1 & 3 \\
2 & 4
\end{pmatrix}
\]

is displayed (using `image`) as:

\[
\begin{pmatrix}
3 & 4 \\
1 & 2
\end{pmatrix}
\]

It is recommended that the source data for this function is a matrix. Replacement of NAs in a data frame extends processing time for this function.

epi.betabuster An R version of Wes Johnson and Chun-Lung Su’s Betabuster

Description

A function to return shape1 and shape2 parameters for a beta distribution, based on expert elicitation.

Usage

epi.betabuster(mode, conf, greaterthan, x, conf.level = 0.95, max.shape1 = 100, step = 0.001)
**Arguments**

- **mode**: scalar, the mode of the variable of interest. Must be a number between 0 and 1.
- **conf**: level of confidence (expressed on a 0 to 1 scale) that the true value of the variable of interest is greater or less than argument $x$.
- **greaterthan**: logical, if TRUE you are making the statement that you are conf confident that the true value of the variable of interest is greater than $x$. If FALSE you are making the statement that you are conf confident that the true value of the variable of interest is less than $x$.
- **x**: scalar, value of the variable of interest (see above).
- **conf.level**: magnitude of the returned confidence interval for the estimated beta distribution. Must be a single number between 0 and 1.
- **max.shape1**: scalar, maximum value of the shape1 parameter for the beta distribution.
- **step**: scalar, step value for the shape1 parameter. See details.

**Details**

The beta distribution has two parameters: shape1 and shape2, corresponding to $a$ and $b$ in the original version of BetaBuster. If $r$ equals the number of times an event has occurred after $n$ trials, shape1 = $(r + 1)$ and shape2 = $(n - r + 1)$.

**Value**

A list containing the following:

- **shape1**: the shape1 parameter for the estimated beta distribution.
- **shape2**: the shape2 parameter for the estimated beta distribution.
- **mode**: the mode of the estimated beta distribution.
- **mean**: the mean of the estimated beta distribution.
- **median**: the median of the estimated beta distribution.
- **lower**: the lower bound of the confidence interval of the estimated beta distribution.
- **upper**: the upper bound of the confidence interval of the estimated beta distribution.
- **variance**: the variance of the estimated beta distribution.

**Author(s)**

Simon Firestone (Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Australia) with acknowledgements to Wes Johnson and Chun-Lung Su for the original standalone software.

**References**

Examples

## EXAMPLE 1:
## If a scientist is asked for their best guess for the diagnostic sensitivity
## of a particular test and the answer is 0.90, and if they are also willing
## to assert that they are 80% certain that the sensitivity is greater than
## 0.75, what are the shape1 and shape2 parameters for a beta distribution
## satisfying these constraints?

rval.beta01 <- epi.betabuster(mode = 0.90, conf = 0.80, greaterthan = TRUE,
     x = 0.75, conf.level = 0.95, max.shape1 = 100, step = 0.001)
rval.beta01$shape1; rval.beta01$shape2

## The shape1 and shape2 parameters for the beta distribution that satisfy the
## constraints listed above are 9.875 and 1.986, respectively.

## This beta distribution reflects the probability distribution obtained if
## there were 9 successes, r:
r <- rval.beta01$shape1 - 1; r

## from 10 trials, n:
n <- rval.beta01$shape2 + rval.beta01$shape1 - 2; n

dat.df01 <- data.frame(x = seq(from = 0, to = 1, by = 0.001),
     y = dbeta(x = seq(from = 0, to = 1, by = 0.001),
     shape1 = rval.beta01$shape1, shape2 = rval.beta01$shape2))

## Density plot of the estimated beta distribution:

## Not run:
library(ggplot2)
ggplot(data = dat.df01, aes(x = x, y = y)) +
  geom_line() +
  scale_x_continuous(name = "Test sensitivity") +
  scale_y_continuous(name = "Density")

## End(Not run)

table

epi.blcm.paras | Number of parameters to be inferred and number of informative priors required for a Bayesian latent class model

Description

Returns the number of unknown parameters to be inferred and the number of informative priors likely to be needed for an identifiable Bayesian latent class model to estimate diagnostic sensitivity and specificity in the absence of a gold standard.

Usage

epi.blcm.paras(ntest.dep = 2, ntest.indep = 1, npop = 2)
Arguments

ntest.dep  scalar, the number of conditionally dependent tests evaluated.
ntest.indep  scalar, the number of conditionally independent tests evaluated.
npop  scalar, the number of populations with a distinct prevalence investigated.

Value

A list containing the following:

df  scalar, the degrees of freedom in the available data.
npars  scalar, the number of unknown parameters to be inferred.
ninf.priors  scalar, the number of informative priors likely to be needed for an identifiable model.

Note

A model may still be useful for inference if it has less informative priors, though cautious interpretation is warranted, typically with a sensitivity analysis of the influence of the priors on the findings.

Author(s)

Simon Firestone and Allison Cheung (Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville Victoria, Australia), Nagendra Singanallur (Australian Centre for Disease Preparedness, Newcomb Victoria, Australia).

Examples

```r
## EXAMPLE 1 --- Two conditionally dependent tests, 1 population:
epi.blcm.paras(ntest.dep = 2, ntest.indep = 0, npop = 1)
```

## This model has 3 degrees of freedom. The model has 7 unknown parameters to be inferred. At least 4 informative priors are required.

```r
## EXAMPLE 2 --- Two conditionally dependent tests, 2 populations:
epi.blcm.paras(ntest.dep = 2, ntest.indep = 0, npop = 2)
```

## This model has 6 degrees of freedom. The model has 8 unknown parameters to be inferred. At least 2 informative priors are required.

```r
## EXAMPLE 3 --- Two conditionally dependent tests, 3 populations:
epi.blcm.paras(ntest.dep = 2, ntest.indep = 0, npop = 3)
```

## This model has 9 degrees of freedom. The model has 9 unknown parameters to be inferred. This model may be able to proceed without informative priors.

```r
## EXAMPLE 4 --- Two conditionally dependent tests, 1 independent test, 1 population:
```
epi.bohning

epi.blcm.paras(ntest.dep = 2, ntest.indep = 1, npop = 1)

## This model has 7 degrees of freedom. The model has 9 unknown parameters to
## be inferred. At least 2 informative priors are required.

## EXAMPLE 5 --- Two conditionally dependent tests, 1 independent test, 2
## populations:
epi.blcm.paras(ntest.dep = 2, ntest.indep = 1, npop = 2)

## This model has 14 degrees of freedom. The model has 10 unknown parameters to
## be inferred. This model may be able to proceed without informative priors.

## EXAMPLE 6 --- Three conditionally dependent tests, 1 population:
epi.blcm.paras(ntest.dep = 3, ntest.indep = 0, npop = 1)

## This model has 7 degrees of freedom. The model has 13 unknown parameters to
## be inferred. At least 6 informative priors are required.

## EXAMPLE 7 --- Three conditionally dependent tests, 2 populations:
epi.blcm.paras(ntest.dep = 3, ntest.indep = 0, npop = 2)

## This model has 14 degrees of freedom. The model has 14 unknown parameters to
## be inferred. This model may be able to proceed without informative priors.

---

epi.bohning

**Bohning’s test for overdispersion of Poisson data**

**Description**

A test for overdispersion of Poisson data.

**Usage**

epi.bohning(obs, exp, alpha = 0.05)

**Arguments**

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>obs</td>
<td>the observed number of cases in each area.</td>
</tr>
<tr>
<td>exp</td>
<td>the expected number of cases in each area.</td>
</tr>
<tr>
<td>alpha</td>
<td>alpha level to be used for the test of significance. Must be a single number between 0 and 1.</td>
</tr>
</tbody>
</table>

**Value**

A data frame with two elements: test.statistic, Bohning’s test statistic and p.value the associated P-value.
References


Examples

```r
## EXAMPLE 1:
data(epi.SClip)
obs <- epi.SClip$cases
pop <- epi.SClip$population
exp <- (sum(obs) / sum(pop)) * pop
epi.bohning(obs, exp, alpha = 0.05)
```

```
## Bohning's test was used to determine if there was statistically significant
## overdispersion in lip cancer cases across 56 Scottish districts for the
## period 1975 to 1980.
## The test statistic was 53.33. The associated P value was <0.01. We reject
## the null hypothesis of no over dispersion and accept the null hypothesis
## concluding that the lip cancer data are over dispersed.
```

epi.ccc

**Concordance correlation coefficient**

Description

Calculates Lin’s (1989, 2000) concordance correlation coefficient for agreement on a continuous measure.

Usage

```r
epi.ccc(x, y, ci = "z-transform", conf.level = 0.95, rep.measure = FALSE,
subjectid)
```

Arguments

- **x**
a vector, representing the first set of measurements.

- **y**
a vector, representing the second set of measurements.

- **ci**
a character string, indicating the method to be used. Options are `z-transform` or `asymptotic`.

- **conf.level**
magnitude of the returned confidence interval. Must be a single number between 0 and 1.

- **rep.measure**
logical. If TRUE there are repeated observations across subject.

- **subjectid**
a factor providing details of the observer identifier if `rep.measure == TRUE`. 
Details

Computes Lin’s (1989, 2000) concordance correlation coefficient for agreement on a continuous measure obtained by two methods. The concordance correlation coefficient combines measures of both precision and accuracy to determine how far the observed data deviate from the line of perfect concordance (that is, the line at 45 degrees on a square scatter plot). Lin’s coefficient increases in value as a function of the nearness of the data’s reduced major axis to the line of perfect concordance (the accuracy of the data) and of the tightness of the data about its reduced major axis (the precision of the data).

Both x and y values need to be present for a measurement pair to be included in the analysis. If either or both values are missing (i.e. coded NA) then the measurement pair is deleted before analysis.

Value

A list containing the following:

- **rho.c**: the concordance correlation coefficient.
- **s.shift**: the scale shift.
- **l.shift**: the location shift.
- **C.b**: a bias correction factor that measures how far the best-fit line deviates from a line at 45 degrees. No deviation from the 45 degree line occurs when C.b = 1. See Lin (1989, page 258).
- **blalt**: a data frame with two columns: mean the mean of each pair of measurements, delta vector y minus vector x.
- **sblalt**: a data frame listing the average difference between the two sets of measurements, the standard deviation of the difference between the two sets of measurements and the lower and upper confidence limits of the difference between the two sets of measurements. If rep.measure == TRUE the confidence interval of the difference is adjusted to account for repeated observations across individual subjects.
- **nmissing**: a count of the number of measurement pairs ignored due to missingness.

References


See Also

epi.occc

Examples

```r
## EXAMPLE 1:
set.seed(seed = 1234)
method1 <- rnorm(n = 100, mean = 0, sd = 1)
method2 <- method1 + runif(n = 100, min = -0.25, max = 0.25)

## Add some missing values:
method1[50] <- NA
method2[75] <- NA
dat.df01 <- data.frame(method1, method2)
rval.ccc01 <- epi.ccc(method1, method2, ci = "z-transform", conf.level = 0.95, rep.measure = FALSE)
rval.lab01 <- data.frame(lab = paste("CCC: ",
round(rval.ccc01$rho.c[,1], digits = 2), " (95% CI ",
round(rval.ccc01$rho.c[,2], digits = 2), " - ",
round(rval.ccc01$rho.c[,3], digits = 2), ")", sep = ""))

z <- lm(method2 ~ method1)
alpha <- summary(z)$coefficients[1,1]
beta <- summary(z)$coefficients[2,1]
rval.lm01 <- data.frame(alpha, beta)
```
## Concordance correlation plot:
## Not run:
library(ggplot2)

ggplot(data = dat.df01, aes(x = method1, y = method2)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1) +
  geom_abline(data = rval.lm01, aes(intercept = alpha, slope = beta),
             linetype = "dashed") +
  scale_x_continuous(limits = c(0,3), name = "Method 1") +
  scale_y_continuous(limits = c(0,3), name = "Method 2") +
  geom_text(data = rval.lab01, x = 0.5, y = 2.95, label = rval.lab01$lab) +
  coord_fixed(ratio = 1 / 1)

## In this plot the dashed line represents the line of perfect concordance.
## The solid line represents the reduced major axis.
## End(Not run)

## EXAMPLE 2:
## Bland and Altman plot (Figure 2 from Bland and Altman 1986):
x <- c(494, 395, 516, 434, 476, 557, 413, 442, 650, 433, 417, 656, 267,
       478, 178, 423, 427)
y <- c(512, 430, 520, 428, 500, 600, 364, 380, 658, 445, 432, 626, 260,
       477, 259, 350, 451)
rval.ccc02 <- epi.ccc(x, y, ci = "z-transform", conf.level = 0.95,
                      rep.measure = FALSE)
rval.df02 <- data.frame(mean = rval.ccc02$blalt[,1],
                        delta = rval.ccc02$blalt[,2])
## Not run:
library(ggplot2)

ggplot(data = rval.ccc02$blalt, aes(x = mean, y = delta)) +
  geom_point() +
  geom_hline(data = rval.ccc02$sblalt, aes(yintercept = lower), linetype = 2) +
  geom_hline(data = rval.ccc02$sblalt, aes(yintercept = upper), linetype = 2) +
  geom_hline(data = rval.ccc02$sblalt, aes(yintercept = est), linetype = 1) +
  scale_x_continuous(limits = c(0,800),
                    name = "Average PEFR by two meters (L/min)") +
  scale_y_continuous(limits = c(-150,150),
                    name = "Difference in PEFR (L/min)")
## End(Not run)

## EXAMPLE 3:
## Setting limits of agreement when your data are skewed. See Euser et al.
## (2008) for details and Rashid et al. (2018) for an applied example.
x <- c(0, 210, 15, 90, 0, 0, 15, 0, 0, 15, 0, 15, 0, 0, 0, 15, 0, 0, 15, 0, 153, 0, 0, 15,
rval.ccc03 <- epi.ccc(x, y, ci = "z-transform", conf.level = 0.95, rep.measure = FALSE)

## Not run:
library(ggplot2)

ggplot(data = rval.ccc03$blalt, aes(x = mean, y = delta)) +
  geom_point() +
  geom_hline(data = rval.ccc03$sblalt, aes(yintercept = lower), linetype = 2) +
  geom_hline(data = rval.ccc03$sblalt, aes(yintercept = upper), linetype = 2) +
  geom_hline(data = rval.ccc03$sblalt, aes(yintercept = est), linetype = 1) +
  scale_x_continuous(limits = c(0,8000), name = "Average of the two measurements") +
  scale_y_continuous(limits = c(-8000,8000), name = "Difference in the two measurements")

## End(Not run)

## In the above plot the spread of the differences increases with increasing
## mean of the observations. The Bland Altman limits of agreement should be
## calculated on a log scale.

logx <- log(x + 50, base = 10)
logy <- log(y + 50, base = 10)

log10.ccc03 <- epi.ccc(x = logx, y = logy, ci = "z-transform", conf.level = 0.95, rep.measure = FALSE)

## Transform the limits of agreement back to the original scale by taking
## anti-logs. If the limits of agreement for Z = log10(x) are between -a
## and +a, with a = 1.96 * s, the ratio between two measures on the original
## scale is between 10^-a and 10^a. See page 979 of Euser et al. (2008).
a <- 1.96 * log10.ccc03$sblalt$delta.sd

## For a given value for the mean Xbar, it can be shown that x - y is between
## -2Xbar(10^a - 1) / (10^a + 1) and +2Xbar(10^a - 1) / (10^a + 1):

Xbar = seq(from = 0, to = 8000, by = 100)
Xbar.low <- (-2 * Xbar * (10^a - 1)) / (10^a + 1)
Xbar.upp <- (+2 * Xbar * (10^a - 1)) / (10^a + 1)

## Not run:
library(ggplot2)
ggplot(data = rval.ccc03$blalt, aes(x = mean, y = delta)) +
  geom_point() +
  geom_line(data = limits, aes(x = mean, y = lower), linetype = 2) +
  geom_line(data = limits, aes(x = mean, y = upper), linetype = 2) +
  geom_line(data = limits, aes(x = mean, y = 0), linetype = 1) +
  scale_x_continuous(limits = c(0,8000),
                     name = "Average of the two measurements") +
  scale_y_continuous(limits = c(-8000,8000),
                     name = "Difference in the two measurements")

## End(Not run)

epi.conf

Confidence intervals for means, proportions, incidence, and standardised mortality ratios

Description
Computes confidence intervals for means, proportions, incidence, and standardised mortality ratios.

Usage
epi.conf(dat, ctype = "mean.single", method, N, design = 1,
          conf.level = 0.95)

Arguments
dat         the data, either a vector or a matrix depending on the method chosen.
ctype       a character string indicating the type of confidence interval to calculate. Options are mean.single, mean.unpaired, mean.paired, prop.single, prop.unpaired, prop.paired, prevalence, inc.risk, inc.rate, odds, ratio and smr.
method      a character string indicating the method to use. Where ctype = "inc.risk" or ctype = "prevalence" options are exact, wilson, fleiss, agresti, clopper-pearson and jeffreys. Where ctype = "inc.rate" options are exact and byar.
N            scalar, representing the population size.
design      scalar, representing the design effect.
conf.level  magnitude of the returned confidence interval. Must be a single number between 0 and 1.

Details
Method mean.single requires a vector as input. Method mean.unpaired requires a two-column data frame; the first column defining the groups must be a factor. Method mean.paired requires a two-column data frame; one column for each group. Method prop.single requires a two-column matrix; the first column specifies the number of positives, the second column specifies the number of negatives. Methods prop.unpaired and prop.paired require a four-column matrix; columns 1 and 2 specify the number of positives and negatives for the first group, columns 3 and 4 specify
the number of positives and negatives for the second group. Method *prevalence* and *inc.risk* require a two-column matrix; the first column specifies the number of positives, the second column specifies the total number tested. Method *inc.rate* requires a two-column matrix; the first column specifies the number of positives, the second column specifies individual time at risk. Method *odds* requires a two-column matrix; the first column specifies the number of positives, the second column specifies the number of negatives. Method *ratio* requires a two-column matrix; the first column specifies the total number of positives, the second column specifies the total number tested.

The methodology implemented here follows Altman, Machin, Bryant, and Gardner (2000). Where method is *inc.risk* or *prevalence* if the numerator equals zero the lower bound of the confidence interval estimate is set to zero. Where method is *smr* the method of Dobson et al. (1991) is used. A summary of the methods used for each of the confidence interval calculations in this function is as follows:

<table>
<thead>
<tr>
<th>ctype-method</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean.single</td>
<td>Altman et al. (2000)</td>
</tr>
<tr>
<td>mean.unpaired</td>
<td>Altman et al. (2000)</td>
</tr>
<tr>
<td>mean.paired</td>
<td>Altman et al. (2000)</td>
</tr>
<tr>
<td>prop.single</td>
<td>Altman et al. (2000)</td>
</tr>
<tr>
<td>prop.unpaired</td>
<td>Altman et al. (2000)</td>
</tr>
<tr>
<td>prop.paired</td>
<td>Altman et al. (2000)</td>
</tr>
<tr>
<td>inc.risk, exact</td>
<td>Collett (1999)</td>
</tr>
<tr>
<td>inc.risk, wilson</td>
<td>Rothman (2002)</td>
</tr>
<tr>
<td>inc.risk, fleiss</td>
<td>Fleiss (1981)</td>
</tr>
<tr>
<td>inc.risk, agresti</td>
<td>Agresti and Coull (1998)</td>
</tr>
<tr>
<td>inc.risk, clopper-pearson</td>
<td>Clopper and Pearson (1934)</td>
</tr>
<tr>
<td>inc.risk, jeffreys</td>
<td>Brown et al. (2001)</td>
</tr>
<tr>
<td>prevalence, exact</td>
<td>Collett (1999)</td>
</tr>
<tr>
<td>prevalence, wilson</td>
<td>Rothman (2002)</td>
</tr>
<tr>
<td>prevalence, fleiss</td>
<td>Fleiss (1981)</td>
</tr>
<tr>
<td>prevalence, agresti</td>
<td>Agresti and Coull (1998)</td>
</tr>
<tr>
<td>prevalence, clopper-pearson</td>
<td>Clopper and Pearson (1934)</td>
</tr>
<tr>
<td>prevalence, jeffreys</td>
<td>Brown et al. (2001)</td>
</tr>
<tr>
<td>inc.rate, exact</td>
<td>Ulm (1990)</td>
</tr>
<tr>
<td>inc.rate, byar</td>
<td>Rothman (2002)</td>
</tr>
<tr>
<td>odds</td>
<td>Ederer and Mantel (1974)</td>
</tr>
<tr>
<td>ratio</td>
<td>Ederer and Mantel (1974)</td>
</tr>
<tr>
<td>smr</td>
<td>Dobson et al. (1991)</td>
</tr>
</tbody>
</table>

The Wald interval often has inadequate coverage, particularly for small sample sizes and proportion estimates close to 0 or 1. Conversely, the Clopper-Pearson Exact method is conservative and tends to produce wider intervals than necessary. Brown et al. recommends the Wilson or Jeffreys methods when sample sizes are small and Agresti-Coull, Wilson, or Jeffreys methods for larger sample sizes.
The Clopper-Pearson interval is an early and very common method for calculating binomial confidence intervals. The Clopper-Pearson interval is sometimes called an 'exact' method, because it is based on the cumulative probabilities of the binomial distribution (i.e. exactly the correct distribution rather than an approximation).

The design effect is used to adjust the confidence interval around a prevalence or incidence risk estimate in the presence of clustering. The design effect is a measure of the variability between clusters and is calculated as the ratio of the variance calculated assuming a complex sample design divided by the variance calculated assuming simple random sampling. Adjustment for the effect of clustering can only be done on those prevalence and incidence risk methods that return a standard error (i.e. method = "wilson" or method = "fleiss").

References


Examples

```r
## EXAMPLE 1:
dat.v01 <- rnorm(n = 100, mean = 0, sd = 1)
epi.conf(dat = dat.v01, ctype = "mean.single")

## EXAMPLE 2:
group <- c(rep("A", times = 5), rep("B", times = 5))
```
val = round(c(rnorm(n = 5, mean = 10, sd = 5),
            rnorm(n = 5, mean = 7, sd = 5)), digits = 0)
dat.df02 <- data.frame(group = group, val = val)
epi.conf(dat = dat.df02, ctype = "mean.unpaired")

## EXAMPLE 3:
## Two paired samples (Altman et al. 2000, page 31):
## Systolic blood pressure levels were measured in 16 middle-aged men
## before and after a standard exercise test. The mean rise in systolic
## blood pressure was 6.6 mmHg. The standard deviation of the difference
## was 6.0 mm Hg. The standard error of the mean difference was 1.49 mm Hg.

after <- c(152,152,134,148,144,134,150,146,174,162,162,146,156,132,126)
dat.df03 <- data.frame(before, after)
epi.conf(dat = dat.df03, ctype = "mean.paired", conf.level = 0.95)

## The 95% confidence interval for the population value of the mean
## systolic blood pressure increase after standard exercise was 3.4 to 9.8
## mm Hg.

## EXAMPLE 4:
## Single sample (Altman et al. 2000, page 47):
## Out of 263 giving their views on the use of personal computers in
## general practice, 81 thought that the privacy of their medical file
## had been reduced.

pos <- 81
neg <- (263 - 81)
dat.m04 <- as.matrix(cbind(pos, neg))
round(epi.conf(dat = dat.m04, ctype = "prop.single"), digits = 3)

## The 95% confidence interval for the population value of the proportion
## of patients thinking their privacy was reduced was from 0.255 to 0.366.

## EXAMPLE 5:
## Two samples, unpaired (Altman et al. 2000, page 49):
## Goodfield et al. report adverse effects in 85 patients receiving either
## terbinafine or placebo treatment for dermatophyte onchomycosis.
## Out of 56 patients receiving terbinafine, 5 patients experienced
## adverse effects. Out of 29 patients receiving a placebo, none experienced
## adverse effects.

grp1 <- matrix(cbind(5, 51), ncol = 2)
grp2 <- matrix(cbind(0, 29), ncol = 2)
dat.m05 <- as.matrix(cbind(grp1, grp2))
round(epi.conf(dat = dat.m05, ctype = "prop.unpaired"), digits = 3)

## The 95% confidence interval for the difference between the two groups is
## from -0.038 to +0.193.
## EXAMPLE 6:
Two samples, paired (Altman et al. 2000, page 53):
In a reliability exercise, 41 patients were randomly selected from those
who had undergone a thalium-201 stress test. The 41 sets of images were
classified as normal or not by the core thalium laboratory and,
independently, by clinical investigators from different centres.
Of the 19 samples identified as ischaemic by clinical investigators
5 were identified as ischaemic by the laboratory. Of the 22 samples
identified as normal by clinical investigators 0 were identified as
ischaemic by the laboratory.

<table>
<thead>
<tr>
<th>Clinic</th>
<th>Laboratory</th>
<th></th>
</tr>
</thead>
</table>
| Ischaemic | 14 | 5 | 19
| Normal | 0 | 22 | 22
| Total | 14 | 27 | 41

```
# epi.conf

dat.m06 <- as.matrix(cbind(14, 5, 0, 22))
round(epi.conf(dat = dat.m06, ctype = "prop.paired", conf.level = 0.95), digits = 3)
```

The 95% confidence interval for the population difference in
proportions is 0.011 to 0.226 or approximately +1% to +23%.

## EXAMPLE 7:
A herd of 1000 cattle were tested for brucellosis. Four samples out of 200
test returned a positive result. Assuming 100% test sensitivity and
specificity, what is the estimated prevalence of brucellosis in this
group of animals?

```
pos <- 4; pop <- 200
dat.m07 <- as.matrix(cbind(pos, pop))
epi.conf(dat = dat.m07, ctype = "prevalence", method = "exact", N = 1000,
design = 1, conf.level = 0.95) * 100
```

The estimated prevalence of brucellosis in this herd is 2.0 (95% CI 0.54 to
5.0) cases per 100 cattle at risk.

## EXAMPLE 8:
The observed disease counts and population size in four areas are provided
below. What are the the standardised morbidity ratios of disease for each
area and their 95% confidence intervals?

```
obs <- c(5, 10, 12, 18); pop <- c(234, 189, 432, 812)
dat.m08 <- as.matrix(cbind(obs, pop))
round(epi.conf(dat = dat.m08, ctype = "smr"), digits = 2)
```
## EXAMPLE 9:

A survey has been conducted to determine the proportion of broilers protected from a given disease following vaccination. We assume that the intra-cluster correlation coefficient for protection (also known as the rate of homogeneity, rho) is 0.4 and the average number of birds per flock is 30. A total of 5898 birds from a total of 10363 were identified as protected. What proportion of birds are protected and what is the 95% confidence interval for this estimate?

Calculate the design effect, given \( \frac{\text{design} - 1}{\bar{n} - 1} \), where \( \bar{n} \) equals the average number of individuals sampled per cluster:

\[
D = 0.4 \times (30 - 1) + 1; D
\]

The design effect is 12.6. Now calculate the proportion protected. We set \( N \) to large number.

\[
\text{dat.m09 <- as.matrix(cbind(5898, 10363))}
\]

\[
\text{epi.conf(dat = dat.m09, ctype = "prevalence", method = "fleiss", N = 1000000, design = D, conf.level = 0.95)}
\]

The estimated proportion of the population protected is 0.57 (95% CI 0.53 to 0.60). Recalculate this estimate assuming the data were from a simple random sample (i.e. where the design effect is one):

\[
\text{epi.conf(dat = dat.m09, ctype = "prevalence", method = "fleiss", N = 1000000, design = 1, conf.level = 0.95)}
\]

If we had mistakenly assumed that data were a simple random sample the confidence interval for the proportion of birds protect would have been 0.56 -- 0.58.

---

epi.convgrid

**Convert British National Grid georeferences to easting and northing coordinates**

**Description**

Convert British National Grid georeferences to British National Grid (EPSG 27700) easting and northing coordinates.

**Usage**

epi.convgrid(osref)
Arguments

**osref**

A vector of character strings listing the British National Grid georeferences to be converted.

Note

If an invalid georeference is encountered in the vector `os.ref` the method returns a `NA`.

Examples

```r
## EXAMPLE 1:
os.ref <- c("SJ505585","SJ488573","SJ652636")
epi.convgrid(os.ref)
```

```r
os.ref <- c("SJ505585","SJ488573","ZZ123456")
epi.convgrid(os.ref)
```

---

**epi.cp**

*Extract unique covariate patterns from a data set*

### Description

Extract the set of unique patterns from a set of covariates (explanatory variables).

### Usage

```r
epi.cp(dat)
```

### Arguments

**dat**

An *i* row by *j* column data frame where the *i* rows represent individual observations and the *m* columns represent a set of *m* covariates. The function allows for one or more covariates for each observation.

### Details

This function extracts the *k* unique covariate patterns in a data set comprised of *i* observations, labelling them from 1 to *k*. The frequency of occurrence of each covariate pattern is listed. A vector of length *i* is also returned, listing the 1:*k* covariate pattern identifier for each observation.
Value

A list containing the following:

cov.pattern  a data frame with columns: id the unique covariate pattern identifier (labelled 1 to k), n the number of occasions each of the listed covariate pattern appears in the data, and the unique covariate combinations.

id  a vector of length i listing the 1:k covariate pattern identifier for each observation.

Author(s)

Thanks to Johann Popp and Mathew Jay for providing code and suggestions to enhance the utility of this function.

References


Examples

## EXAMPLE 1:

### Generate a set of covariates:

set.seed(seed = 1234)
obs <- round(runif(n = 100, min = 0, max = 1), digits = 0)
v1 <- round(runif(n = 100, min = 0, max = 4), digits = 0)
v2 <- round(runif(n = 100, min = 0, max = 4), digits = 0)
dat.df01 <- data.frame(obs, v1, v2)
dat.glm01 <- glm(obs ~ v1 + v2, family = binomial, data = dat.df01)
dat.mf01 <- model.frame(dat.glm01)

### Covariate pattern. Drop the first column of dat.mf01 (since column 1 is the outcome variable):

epi.cp(dat.mf01[,2:3])

### There are 25 covariate patterns in this data set. Subject 100 has covariate pattern 21.

---

epi.cpresids  Covariate pattern residuals from a logistic regression model

Description

Returns covariate pattern residuals and delta betas from a logistic regression model.
Usage

epi.cpresids(obs, fit, covpattern)

Arguments

obs a vector of observed values (i.e. counts of ‘successes’) for each covariate pattern.

fit a vector defining the predicted (i.e. fitted) probability of success for each covariate pattern.

covpattern a epi.cp object.

Value

A data frame with 13 elements: cpid the covariate pattern identifier, n the number of subjects in this covariate pattern, obs the observed number of successes, pred the predicted number of successes, raw the raw residuals, sraw the standardised raw residuals, pearson the Pearson residuals, spearson the standardised Pearson residuals, deviance the deviance residuals, leverage leverage, deltabeta the delta-betas, sdeltabeta the standardised delta-betas, and deltachi delta chi statistics.

References


See Also

epi.cp

Examples

## EXAMPLE 1:
dat.glm01 <- glm(case ~ spontaneous + induced, family = binomial(),
                  data = infert)

## Covariate patterns:
dat.mf01 <- model.frame(dat.glm01)
dat.cp01 <- epi.cp(dat.mf01[-1])

dat.obs01 <- as.vector(by(infert$case, as.factor(dat.cp01$id),
                         FUN = sum))
dat.fit01 <- as.vector(by(fitted(dat.glm01), as.factor(dat.cp01$id),
                        FUN = min))
dat.cpr01 <- epi.cpresids(obs = dat.obs01, fit = dat.fit01,
                          covpattern = dat.cp01)
head(dat.cpr01)
Description

Computes descriptive statistics for a numeric vector or a table of frequencies for a factor.

Usage

epi.descriptives(dat, conf.level = 0.95)

Arguments

dat either a numeric vector or a factor.
conf.level magnitude of the returned confidence intervals. Must be a single number between 0 and 1.

Value

If dat is numeric a list containing the following:

- arithmetic n number of observations, mean arithmetic mean, sd arithmetic standard deviation, q25 25th quantile, q50 50th quantile, q75 75th quantile, lower lower bound of the confidence interval, upper upper bound of the confidence interval, min minimum value, max maximum value, and na number of missing values.
- geometric n number of observations, mean geometric mean, sd geometric standard deviation, q25 25th quantile, q50 50th quantile, q75 75th quantile, lower lower bound of the confidence interval, upper upper bound of the confidence interval, min minimum value, max maximum value, and na number of missing values.
- symmetry skewness and kurtosis.

If dat is a factor a data frame listing:

- level The levels of the factor
- n The frequency of the respective factor level, including the column totals.

Examples

## EXAMPLE 1:
## Generate some data:
id <- 1:100
n <- rnorm(100, mean = 0, sd = 1)
dat.df01 <- data.frame(id, n)

# Add missing values:
missing <- dat.df01$id %in% sample(dat.df01$id, size = 20)
dat.df01$n[missing] <- NA
## EXAMPLE 2:
## Generate some data:
## n <- 1000; p.exp <- 0.50; p.dis <- 0.75
## strata <- c(rep("A", times = n / 2), rep("B", times = n / 2))
## exp <- rbinom(n = n, size = 1, prob = p.exp)
## dis <- rbinom(n = n, size = 1, prob = p.dis)
## dat.df02 <- data.frame(strata, exp, dis)
## dat.df02$strata <- factor(dat.df02$strata)
## dat.df02$exp <- factor(dat.df02$exp, levels = c("1", "0"))
## head(dat.df02)
## epi.descriptives(dat.df02$exp, conf.level = 0.95)

---

**epi.dgamma**

Estimate the precision of a [structured] heterogeneity term

**Description**

Returns the precision of a [structured] heterogeneity term after one has specified the amount of variation a priori.

**Usage**

```r
epi.dgamma(rr, quantiles = c(0.05, 0.95))
```

**Arguments**

- **rr**
  - the lower and upper limits of relative risk, estimated *a priori*.
- **quantiles**
  - a vector of length two defining the quantiles of the lower and upper relative risk estimates.

**Value**

Returns the precision (the inverse variance) of the heterogeneity term.

**References**

Best, NG. WinBUGS 1.3.1 Short Course, Brisbane, November 2000.
Examples

```r
## EXAMPLE 1:
## Suppose we are expecting the lower 5% and upper 95% confidence interval
## of relative risk in a data set to be 0.5 and 3.0, respectively.
## A prior estimate of the precision of the heterogeneity term would be:

tau <- epi.dgamma(rr = c(0.5, 3.0), quantiles = c(0.05, 0.95))

tau
```

```
## The estimate of the precision of the heterogeneity term (tau) is 3.37.
## This can be re-expressed using the gamma distribution. We set the mean of the
## distribution as tau and specify a large variance (that is, we are not
## certain about tau).

mean <- tau; var <- 1000
shape <- mean^2 / var
inv.scale <- mean / var

## In WinBUGS the precision of the heterogeneity term is parameterised
## as tau ~ dgamma(shape, inv.scale). Plot the probability density function
## of tau:

z <- seq(0.01, 10, by = 0.01)
fz <- dgamma(z, shape = shape, scale = 1 / inv.scale)
plot(x = z, y = fz, type = "l", ylab = "Probability density of tau")
```

---

epi.directadj  

**Directly adjusted incidence rate estimates**

Description

Compute directly adjusted incidence rate estimates.

Usage

```r
epi.directadj(obs, tar, std, units = 1, conf.level = 0.95)
```

Arguments

- **obs**: a matrix representing the observed number of events. Rows represent strata (e.g. region); columns represent the variables to be adjusted for (e.g. age class, gender). The sum of each row will equal the total number of events for each stratum. The rows of the obs matrix must be named with the appropriate strata names and the columns of obs must be named with the appropriate level identifiers for each explanatory variable. See the example, below.
epi.directadj

tar a matrix representing population time at risk. Rows represent strata (e.g. region); columns represent the variables to be adjusted for (e.g. age class, gender). The sum of each row will equal the total population time at risk for each stratum. The rows of the pop matrix must be named with the appropriate strata names and the columns of pop must be named with the appropriate level identifiers for each explanatory variable. See the example, below.

std a matrix representing the standard population size for the different levels of the covariate to be adjusted for. The columns of std must be named with the appropriate level identifiers for each explanatory variable.

units multiplier for the incidence rate estimates.

conf.level magnitude of the returned confidence interval. Must be a single number between 0 and 1.

Details

This function returns unadjusted (crude) and directly adjusted incidence rate estimates for each of the specified population strata. The term ‘covariate’ is used here to refer to the factors we want to control (i.e. adjust) for when calculating the directly adjusted incidence rate estimates.

When the outcome of interest is rare, the confidence intervals for the adjusted incidence rates returned by this function (based on Fay and Feuer, 1997) will be appropriate for incidence risk data. In this situation the argument tar is assumed to represent the size of the population at risk (instead of population time at risk). Example 3 (below) provides an approach if you are working with incidence risk data and the outcome of interest is not rare.

Value

A list containing the following:

- crude the crude incidence rate estimates for each stratum-covariate combination.
- crude.strata the crude incidence rate estimates for each stratum.
- adj.strata the directly adjusted incidence rate estimates for each stratum.

Author(s)

Thanks to Karl Ove Hufthammer for helpful suggestions to improve the execution and documentation of this function.

References


See Also

epi.indirectadj

Examples

## EXAMPLE 1 (from Thrusfield 2007 pp. 63 - 64):
## A study was conducted to estimate the seroprevalence of leptospirosis in dogs in Glasgow and Edinburgh, Scotland. Data frame dat.df lists counts of leptospirosis cases and the number of dog years at risk for male and female dogs:

dat.df01 <- data.frame(obs = c(15,46,53,16), tar = c(48,212,180,71),
                  sex = c("M","F","M","F"), city = c("ED","ED","GL","GL"))

obs01 <- matrix(dat.df01$obs, nrow = 2, byrow = TRUE, dimnames = list(c("ED","GL"), c("M","F")))
tar01 <- matrix(dat.df01$tar, nrow = 2, byrow = TRUE, dimnames = list(c("ED","GL"), c("M","F")))

## Create a standard population with equal numbers of male and female dogs:

std01 <- matrix(data = c(250,250), nrow = 1, byrow = TRUE, dimnames = list("", c("M","F")))

## Directly adjusted incidence rates:
epi.directadj(obs01, tar01, std01, units = 1, conf.level = 0.95)

## $crude
## strata cov obs tar est lower upper
## ED M 15 48 0.3125000 0.1749039 0.5154212
## GL M 53 180 0.2944444 0.2205591 0.3851406
## ED F 46 212 0.2169811 0.1588575 0.2894224
## GL F 16 71 0.2253521 0.1288082 0.3659577

## $crude.strata
## strata obs tar est lower upper
## ED 61 250 0.3125000 0.1749039 0.5154212
## GL 69 250 0.2944444 0.2205591 0.3851406

## $adj.strata
## strata obs tar est lower upper
## ED 61 250 0.3125000 0.1749039 0.5154212
## GL 69 250 0.2944444 0.2205591 0.3851406

## The adjusted incidence rate of leptospirosis in Glasgow dogs is 0.31 (95% CI 0.20 to 0.34) cases per 100 dog-years at risk. The confounding effect of gender has been removed by the adjusted incidence rate estimates.
## EXAMPLE 2:
Here we provide a more flexible approach for calculating adjusted incidence rate estimates using Poisson regression. See Frome and Checkoway (1985) for details.

```r
dat(glm02 <- glm(obs ~ city, offset = log(tar), family = poisson, data = dat.df01)
summary(dat glm02)
```

## To obtain adjusted incidence rate estimates, use the predict method on a new data set with the time at risk (tar) variable set to 1 (which means log(tar) = 0). This will return the predicted number of cases per one unit of individual time, i.e. the incidence rate.

```r
dat.pred02 <- predict(object = dat glm02, newdata = data.frame(city = c("ED","GL"), tar = c(1,1)), type = "link", se = TRUE)
conf.level <- 0.95
critval <- qnorm(p = 1 - ((1 - conf.level) / 2), mean = 0, sd = 1)
est <- dat glm02$family$linkinv(dat.pred02$fit)
lower <- dat glm02$family$linkinv(dat.pred02$fit - (critval * dat.pred02$se.fit))
upper <- dat glm02$family$linkinv(dat.pred02$fit + (critval * dat.pred02$se.fit))
round(x = data.frame(est, lower, upper), digits = 3)
```

## Results identical to the crude incidence rate estimates from epi.directadj.

## EXAMPLE 3:
Now adjust for the effect of gender and city and report the adjusted incidence rate estimates for each city:

```r
dat glm03 <- glm(obs ~ city + sex, offset = log(tar), family = poisson, data = dat df01)
dat.pred03 <- predict(object = dat glm03, newdata = data.frame(sex = c("F","F"), city = c("ED","GL"), tar = c(1,1)), type = "link", se.fit = TRUE)
conf.level <- 0.95
critval <- qnorm(p = 1 - ((1 - conf.level) / 2), mean = 0, sd = 1)
est <- dat glm03$family$linkinv(dat.pred03$fit)
lower <- dat glm03$family$linkinv(dat.pred03$fit - (critval * dat.pred03$se.fit))
upper <- dat glm03$family$linkinv(dat.pred03$fit + (critval * dat.pred03$se.fit))
round(x = data.frame(est, lower, upper), digits = 3)
```
## Using Poisson regression the gender adjusted incidence rate of leptospirosis in Glasgow dogs was 22 (95% CI 15 to 32) cases per 100 dog-years at risk. These results won't be the same as those using direct adjustment because for direct adjustment we use a contrived standard population.

## EXAMPLE 4 --- Logistic regression to return adjusted incidence risk estimates:

Say, for argument's sake, that we are now working with incidence risk data. Here we'll re-label the variable 'tar' (time at risk) as 'pop' (population size). We adjust for the effect of gender and city and report the adjusted incidence risk of canine leptospirosis estimates for each city:

```r
dat.df01$pop <- dat.df01$tar
dat.glm04 <- glm(cbind(obs, pop - obs) ~ city + sex, family = "binomial", data = dat.df01) dat.pred04 <- predict(object = dat.glm04, newdata = data.frame(sex = c("F","F"), city = c("ED","GL")), type = "link", se.fit = TRUE) conf.level <- 0.95 critval <- qnorm(p = 1 - ((1 - conf.level) / 2), mean = 0, sd = 1) est <- dat.glm04$family$linkinv(dat.pred04$fit) lower <- dat.glm04$family$linkinv(dat.pred04$fit - (critval * dat.pred04$se.fit)) upper <- dat.glm04$family$linkinv(dat.pred04$fit + (critval * dat.pred04$se.fit)) round(x = data.frame(est, lower, upper), digits = 3)
```

## The adjusted incidence risk of leptospirosis in Glasgow dogs is 22 (95% CI 15 to 30) cases per 100 dogs at risk.

### Description

Converts decimal degrees to degrees, minutes and seconds. Converts degrees, minutes and seconds to decimal degrees.
Usage
epi.dms(dat)

Arguments
dat the data. A one-column matrix is assumed when converting decimal degrees to
degrees, minutes, and seconds. A two-column matrix is assumed when convert-
ing degrees and decimal minutes to decimal degrees. A three-column matrix is
assumed when converting degrees, minutes and seconds to decimal degrees.

Examples

## EXAMPLE 1:
## Degrees, minutes, seconds to decimal degrees:
dat.m01 <- matrix(c(41, 38, 7.836, -40, 40, 27.921),
                   byrow = TRUE, nrow = 2)
epi.dms(dat.m01)

## EXAMPLE 2:
## Decimal degrees to degrees, minutes, seconds:
dat.m02 <- matrix(c(41.63551, -40.67442), nrow = 2)
epi.dms(dat.m02)

description
Mixed-effects meta-analysis of binary outcomes using the DerSimonian and Laird method

Description
 Computes individual study odds or risk ratios for binary outcome data. Computes the summary
odds or risk ratio using the DerSimonian and Laird method. Performs a test of heterogeneity among
trials. Performs a test for the overall difference between groups (that is, after pooling the studies,
do treated groups differ significantly from controls?).

Usage
epi.dsl(ev.trt, n.trt, ev.ctrl, n.ctrl, names, method = "odds.ratio",
        alternative = c("two.sided", "less", "greater"), conf.level = 0.95)

Arguments
ev.trt observed number of events in the treatment group.
n.trt number in the treatment group.
ev.ctrl observed number of events in the control group.
n.ctrl number in the control group.
names character string identifying each trial.

method a character string indicating the method to be used. Options are odds.ratio or risk.ratio.

alternative a character string specifying the alternative hypothesis, must be one of two.sided, greater or less.

conf.level magnitude of the returned confidence interval. Must be a single number between 0 and 1.

Details

alternative = "greater" tests the hypothesis that the DerSimonian and Laird summary measure of association is greater than 1.

Value

A list containing the following:

- OR the odds ratio for each trial and the lower and upper bounds of the confidence interval of the odds ratio for each trial.
- RR the risk ratio for each trial and the lower and upper bounds of the confidence interval of the risk ratio for each trial.
- OR.summary the DerSimonian and Laird summary odds ratio and the lower and upper bounds of the confidence interval of the DerSimonian and Laird summary odds ratio.
- RR.summary the DerSimonian and Laird summary risk ratio and the lower and upper bounds of the confidence interval of the DerSimonian and Laird summary risk ratio.
- weights the inverse variance and DerSimonian and Laird weights for each trial.
- heterogeneity a vector containing Q the heterogeneity test statistic, df the degrees of freedom and its associated P-value.
- Hsq the relative excess of the heterogeneity test statistic Q over the degrees of freedom df.
- Isq the percentage of total variation in study estimates that is due to heterogeneity rather than chance.
- tau.sq the variance of the treatment effect among trials.
- effect a vector containing z the test statistic for overall treatment effect and its associated P-value.

Note

Under the random-effects model, the assumption of a common treatment effect is relaxed, and the effect sizes are assumed to have a normal distribution with variance tau.sq.

Using this method, the DerSimonian and Laird weights are used to compute the pooled odds ratio. The function checks each strata for cells with zero frequencies. If a zero frequency is found in any cell, 0.5 is added to all cells within the strata.
References


See Also

epi.iv, epi.mh, epi.smd

Examples

```r
## EXAMPLE 1:
data(epi.epidural)
epi.dsl(ev.trt = epi.epidural$ev.trt, n.trt = epi.epidural$n.trt,
ev.ctrl = epi.epidural$ev.ctrl, n.ctrl = epi.epidural$n.ctrl,
names = as.character(epi.epidural$trial), method = "odds.ratio",
alternative = "two.sided", conf.level = 0.95)
```

---

**epi.edr**  
*Estimated dissemination ratio*

**Description**

Computes estimated dissemination ratios on the basis of a vector of count data (usually incident cases identified on each day of an epidemic).

**Usage**

```r
epi.edr(dat, n = 4, conf.level = 0.95, nsim = 99, na.zero = TRUE)
```

**Arguments**

- `dat`  
  a numeric vector listing the number of incident cases for each day of an epidemic.

- `n`  
  scalar, defining the number of days to be used when computing the estimated dissemination ratio.

- `conf.level`  
  magnitude of the returned confidence interval. Must be a single number between 0 and 1.

- `nsim`  
  scalar, defining the number of simulations to be used for the confidence interval calculations.

- `na.zero`  
  logical, replace NaN or Inf values with zeros?
Details

In infectious disease epidemics the $n$-day estimated dissemination ratio (EDR) at day $i$ equals the total number of incident cases between day $i$ and day $[i-(n-1)]$ (inclusive) divided by the total number of incident cases between day $(i-n)$ and day $(i-2n)+1$ (inclusive). EDR values are often calculated for each day of an epidemic and presented as a time series analysis. If the EDR is consistently less than unity, the epidemic is said to be 'under control'.

A simulation approach is used to calculate confidence intervals around each daily EDR estimate. The numerator and denominator of the EDR estimate for each day is taken in turn and a random number drawn from a Poisson distribution, using the calculated numerator and denominator value as the mean. EDR is then calculated for these simulated values and the process repeated $\text{nsim}$ times. Confidence intervals are then derived from the vector of simulated values for each day.

Value

Returns the point estimate of the EDR and the lower and upper bounds of the confidence interval of the EDR.

References


Examples

```r
## EXAMPLE 1:
set.seed(1234)
dat <- rpois(n = 50, lambda = 2)
dat.edr01 <- epi.edr(dat, n = 4, conf.level = 0.95, nsim = 99, na.zero = TRUE)
sdate <- as.Date(x = "31/12/2015", format = "%d/%m/%Y") + 1:50
dat.df01 <- data.frame(sdate = sdate, est = dat.edr01$est,
                        low = dat.edr01$lower, upp = dat.edr01$upper)

## Line plot of EDR (and its 95% confidence interval) as a function of calendar time:
## Not run:
library(ggplot2); library(scales)
ggplot() + geom_line(data = dat.df01, aes(x = sdate, y = est)) +
  geom_line(data = dat.df01, aes(x = sdate, y = upp), lty = 3, size = 0.5) +
```

```r
## Not run:
library(ggplot2); library(scales)
ggplot() + geom_line(data = dat.df01, aes(x = sdate, y = est)) +
  geom_line(data = dat.df01, aes(x = sdate, y = upp), lty = 3, size = 0.5) +
```

```r
## Not run:
library(ggplot2); library(scales)
ggplot() + geom_line(data = dat.df01, aes(x = sdate, y = est)) +
  geom_line(data = dat.df01, aes(x = sdate, y = upp), lty = 3, size = 0.5) +
```
The gamma distribution is parameterised in terms of shape ($\alpha$) and scale ($\nu$) parameters. The mean of a given gamma distribution equals $\nu/\alpha$. The variance equals $\nu/\alpha^2$. The empirical Bayes estimate of event risk in each unit of interest equals $(\text{obs} + \nu)/(\text{pop} + \alpha)$.

This technique performs poorly when your data contains large numbers of zero event counts. In this situation a Bayesian approach for estimating $\alpha$ and $\nu$ would be advised.

Value

A data frame with four elements: gamma the mean event risk across all units, phi the variance of event risk across all units, alpha the estimated shape parameter of the gamma distribution, and nu the estimated scale parameter of the gamma distribution.

References


Examples

```r
## EXAMPLE 1:
data(epi.SClip)
obs <- epi.SClip$cases; pop <- epi.SClip$population
est <- epi.empbayes(obs, pop)
crude.p <- ((obs) / (pop)) * 100000
crude.r <- rank(crude.p)
ebay.p <- ((obs + est[4]) / (pop + est[3])) * 100000
dat.df01 <- data.frame(rank = c(crude.r, crude.r),
Method = c(rep("Crude", times = length(crude.r)),
rep("Empirical Bayes", times = length(crude.r)),
est = c(crude.p, ebay.p))

## Scatter plot showing the crude and empirical Bayes adjusted lip cancer
## incidence rates as a function of district rank for the crude lip
## cancer incidence rates:
## Not run:
library(ggplot2)
ggplot(dat = dat.df01, aes(x = rank, y = est, colour = Method)) +
  geom_point() +
  scale_x_continuous(name = "District rank",
   breaks = seq(from = 0, to = 60, by = 10),
   labels = seq(from = 0, to = 60, by = 10),
   limits = c(0,60)) +
  scale_y_continuous(limits = c(0,30), name = "Lip cancer incidence rates
   (cases per 100,000 person years)"
)
## End(Not run)
```

---

**epi.epidural**  
*Rates of use of epidural anaesthesia in trials of caregiver support*

**Description**

This data set provides results of six trials investigating rates of use of epidural anaesthesia during childbirth. Each trial is made up of a group where a caregiver (midwife, nurse) provided support intervention and a group where standard care was provided. The objective was to determine if there were higher rates of epidural use when a caregiver was present at birth.

**Usage**

```r
data(epi.epidural)
```
Format

A data frame with 6 observations on the following 5 variables.

- **trial**  the name and year of the trial.
- **ev.trt** number of births in the caregiver group where an epidural was used.
- **n.trt** number of births in the caregiver group.
- **ev.ctrl** number of births in the standard care group where an epidural was used.
- **n.ctrl** number of births in the standard care group.

References


---

**epi.herdtest**

*Estimate the characteristics of diagnostic tests applied at the herd (group) level*

**Description**

When tests are applied to individuals within a group we may wish to designate the group as being either diseased or non-diseased on the basis of the individual test results. This function estimates sensitivity and specificity of this testing regime at the group (or herd) level.

**Usage**

`epi.herdtest(se, sp, P, N, n, k)`

**Arguments**

- **se**  a vector of length one defining the sensitivity of the individual test used.
- **sp**  a vector of length one defining the specificity of the individual test used.
- **P**  scalar, defining the estimated true prevalence.
- **N**  scalar, defining the herd size.
- **n**  scalar, defining the number of individuals to be tested per group (or herd).
- **k**  scalar, defining the critical number of individuals testing positive that will denote the group as test positive.

**Value**

A data frame with four elements: APpos the probability of obtaining a positive test, APneg the probability of obtaining a negative test, HS\text{e} the estimated group (herd) sensitivity, and HS\text{p} the estimated group (herd) specificity.
Note

The method implemented in this function is based on the hypergeometric distribution.

Author(s)

Ron Thornton, MAF New Zealand, PO Box 2526 Wellington, New Zealand.

References


Examples

```
## EXAMPLE 1:
## We want to estimate the herd-level sensitivity and specificity of
## a testing regime using an individual animal test of sensitivity 0.391
## and specificity 0.964. The estimated true prevalence of disease is 0.12.
## Assume that 60 individuals will be tested per herd and we have
## specified that two or more positive test results identify the herd
## as positive.

epi.herdtest(se = 0.391, sp = 0.964, P = 0.12, N = 1E06, n = 60, k = 2)
```

```
## This testing regime gives a herd sensitivity of 0.95 and a herd
## specificity of 0.36. With a herd sensitivity of 0.95 we can be
## confident that we will declare a herd infected if it is infected.
## With a herd specificity of only 0.36, we will declare 0.64 of disease
## negative herds as infected, so false positives are a problem.
```

Description

Between 1972 and 1980 an industrial waste incinerator operated at a site about 2 kilometres south-west of the town of Coppull in Lancashire, England. Addressing community concerns that there were greater than expected numbers of laryngeal cancer cases in close proximity to the incinerator Diggle et al. (1990) conducted a study investigating risks for laryngeal cancer, using recorded cases of lung cancer as controls. The study area is 20 km x 20 km in size and includes location of residence of patients diagnosed with each cancer type from 1974 to 1983. The site of the incinerator was at easting 354500 and northing 413600.

Usage

```
data(epi.incin)
```
**Format**

A data frame with 974 observations on the following 3 variables.

- **xcoord** easting coordinate (in metres) of each residence.
- **ycoord** northin coordinate (in metres) of each residence.
- **status** disease status: 0 = lung cancer, 1 = laryngeal cancer.

**Source**


**References**


---

### epi.indirectadj

*Indirectly adjusted incidence risk estimates*

**Description**

Compute indirectly adjusted incidence risks and standardised mortality (incidence) ratios.

**Usage**

epi.indirectadj(obs, pop, std, units, conf.level = 0.95)

**Arguments**

- **obs**
  
a one column matrix representing the number of observed number of events in each strata. The dimensions of obs must be named (see the examples, below).

- **pop**
  
a matrix representing population size. Rows represent strata (e.g. region); columns represent the levels of the explanatory variable to be adjusted for (e.g. age class, gender). The sum of each row will equal the total population size within each stratum. If there are no covariates pop will be a one column matrix. The dimensions of the pop matrix must be named (see the examples, below).
std  a one row matrix specifying the standard incidence risks to be applied to each level of the covariate to be adjusted for. The length of std should be one plus the number of covariates to be adjusted for (the additional value represents the incidence risk in the entire population). If there are no explanatory variables to adjust for std is a single number representing the incidence risk in the entire population.

units  multiplier for the incidence risk estimates.

conf.level  magnitude of the returned confidence interval. Must be a single number between 0 and 1.

Details

Indirect standardisation can be performed whenever the stratum-specific incidence risk estimates are either unknown or unreliable. If the stratum-specific incidence risk estimates are known, direct standardisation is preferred.

Confidence intervals for the standardised mortality ratio estimates are based on the Poisson distribution (see Breslow and Day 1987, p 69 - 71 for details).

Value

A list containing the following:

- crude.strata  the crude incidence risk estimates for each stratum.
- adj.strata  the indirectly adjusted incidence risk estimates for each stratum.
- smr  the standardised mortality (incidence) ratios for each stratum.

Author(s)

Thanks to Dr. Telmo Nunes (UISEE/DETSA, Faculdade de Medicina Veterinaria - UTL, Rua Prof. Cid dos Santos, 1300-477 Lisboa Portugal) for details and code for the confidence interval calculations.

References


See Also

epi.directadj
## EXAMPLE 1 (without covariates):
## Adapted from Dohoo, Martin and Stryhn (2009). In this example the frequency
## of tuberculosis is expressed as incidence risk (i.e. the number of
## tuberculosis positive herds divided by the size of the herd population at
## risk). In their text Dohoo et al. present the data as incidence rate (the
## number of tuberculosis positive herds per herd-year at risk).

## Data have been collected on the incidence of tuberculosis in two
## areas ("A" and "B"). Provided are the counts of (new) incident cases and
## counts of the herd population at risk. The standard incidence risk for
## the total population is 0.060 (6 cases per 100 herds at risk):

```r
epis.indirectadj(obs = obs.m01, pop = pop.m01, std = std.m01, units = 100,
conf.level = 0.95)
```

## EXAMPLE 2 (with covariates):
## We now have, for each area, data stratified by herd type (dairy, beef).
## The standard incidence risks for beef herds, dairy herds, and the total
## population are 0.025, 0.085, and 0.060 cases per herd, respectively:

```r
epis.indirectadj(obs = obs.m02, pop = pop.m02, std = std.m02, units = 100,
conf.level = 0.95)
```

```r
# > $crude.strata
# >   est   lower  upper
# > A  5.8   4.404183  7.497845
# > B  6.5   5.430733  7.718222

# > $adj.strata
# >   est   lower  upper
# > A  6.692308  5.076923  8.423077
# > B  5.571429  4.628571  6.557143

# > $smr.strata
# >   obs exp   est   lower  upper
# > A  58 251  1.1153846  0.8461538  1.403846
# > B 138 140  0.9285714  0.7714286  1.092857
```
The crude incidence risk of tuberculosis in area A was 5.8 (95% CI 4.0 to 7.5) cases per 100 herds at risk. The crude incidence risk of tuberculosis in area B was 6.5 (95% CI 5.4 to 7.7) cases per 100 herds at risk.

The indirectly adjusted incidence risk of tuberculosis in area A was 6.7 (95% CI 5.1 to 8.4) cases per 100 herds at risk. The indirectly adjusted incidence risk of tuberculosis in area B was 5.6 (95% CI 4.6 to 6.6) cases per 100 herds at risk.

---

**epi.insthaz**

*Event instantaneous hazard based on Kaplan-Meier survival estimates*

**Description**

Compute event instantaneous hazard on the basis of a Kaplan-Meier survival function.

**Usage**

```r
epi.insthaz(survfit.obj, conf.level = 0.95)
```

**Arguments**

- `survfit.obj`: a survfit object, computed using the survival package.
- `conf.level`: magnitude of the returned confidence interval. Must be a single number between 0 and 1.

**Details**

Computes the instantaneous hazard of the event of interest, equivalent to the proportion of the population failing per unit time.

**Value**

A data frame with three or four elements: `strata` the strata identifier, `time` the observed failure times, `sest` the observed Kaplan-Meier survival function, `slow` the lower bound of the confidence interval for the observed Kaplan-Meier survival function, `supp` the upper bound of the confidence interval for the observed Kaplan-Meier survival function, `hest` the observed instantaneous hazard (the proportion of the population at risk experiencing the event of interest per unit time), `hlow` the lower bound of the confidence interval for the observed instantaneous hazard, and `hupp` the upper bound of the confidence interval for the observed instantaneous hazard.

**References**


Examples

## EXAMPLE 1:
library(survival)
dat.df01 <- lung
dat.df01$status <- ifelse(dat.df01$status == 1, 0, dat.df01$status)
dat.df01$status <- ifelse(dat.df01$status == 2, 1, dat.df01$status)
dat.df01$sex <- factor(dat.df01$sex, levels = c(1,2),
   labels = c("Male","Female"))
lung.km01 <- survfit(Surv(time = time, event = status) ~ 1, data = dat.df01)
lung.haz01 <- epi.insthaz(lung.km01, conf.level = 0.95)
lung.shaz01 <- data.frame(
   time = lowess(lung.haz01$time, lung.haz01$hlow, f = 0.20)$x,
   hest = lowess(lung.haz01$time, lung.haz01$hest, f = 0.20)$y,
   hlow = lowess(lung.haz01$time, lung.haz01$hlow, f = 0.20)$y,
   hupp = lowess(lung.haz01$time, lung.haz01$hupp, f = 0.20)$y)
plot(x = lung.haz01$time, y = lung.haz01$hest, xlab = "Time (days)",
   ylab = "Daily probability of event", type = "s",
   col = "grey", ylim = c(0, 0.05))
lines(x = lung.shaz01$time, y = lung.shaz01$hest,
   lty = 1, lwd = 2, col = "black")
lines(x = lung.shaz01$time, y = lung.shaz01$hlow,
   lty = 2, lwd = 1, col = "black")
lines(x = lung.shaz01$time, y = lung.shaz01$hupp,
   lty = 2, lwd = 1, col = "black")

## Not run:
library(ggplot2)

## EXAMPLE 2:
## Now stratify by gender:

## Not run:
library(ggplot2)
ggplot() +
   theme_bw() +
   geom_step(data = lung.haz01, aes(x = time, y = hest), colour = "grey") +
   geom_smooth(data = lung.haz01, aes(x = time, y = hest), method = "loess",
      colour = "black", size = 0.75, linetype = "solid",
      se = FALSE, span = 0.20) +
   geom_smooth(data = lung.haz01, aes(x = time, y = hlow), method = "loess",
      colour = "black", size = 0.5, linetype = "dashed",
      se = FALSE, span = 0.20) +
   geom_smooth(data = lung.haz01, aes(x = time, y = hupp), method = "loess",
      colour = "black", size = 0.5, linetype = "dashed",
      se = FALSE, span = 0.20) +
   scale_x_continuous(limits = c(0,1000), name = "Time (days)") +
   scale_y_continuous(limits = c(0,0.05), name = "Daily probability of event")

## End(Not run)
epi.interaction

Relative excess risk due to interaction in a case-control study

Description

For two binary explanatory variables included in a logistic regression as an interaction term, computes the relative excess risk due to interaction, the proportion of outcomes among those with both exposures attributable to interaction, and the synergy index. Confidence interval calculations are based on the delta method described by Hosmer and Lemeshow (1992).

Usage

epi.interaction(model, coef, param = c("product", "dummy"),
                 type = c("RERI", "APAB", "S"), conf.level = 0.95)

Arguments

model an object of class glm, clogit or coxph.
coef a vector listing the positions of the coefficients of the interaction terms in the model. What row numbers of the regression table summary list the coefficients for the interaction terms included in the model?
epi.interaction

param character string specifying the type of coding used for the variables included in the interaction term. Options are product where two risk factors and one product term are used to represent the interaction and dummy where the two risk factors are combined into a single variable with four levels. See the examples, below, for details.

type character string specifying the type of analysis to be run. Options are RERI the relative excess risk due to interaction, AP[AB] the proportion of disease among those with both exposures that is attributable to interaction of the two exposures, and S the synergy index.

conf.level magnitude of the returned confidence interval. Must be a single number between 0 and 1.

Details

Interaction on an additive scale means that the combined effect of two exposures is greater (or less) than the sum of the individual effects of two exposures. Interaction on a multiplicative scale means that the combined effect of the two exposures is greater (or less) than the product of the individual effects of the two exposures.

This function calculates three indices to assess the presence of additive interaction, as defined by Rothman (1998): (1) the relative excess risk due to interaction (RERI, sometimes called the interaction contrast ratio), (2) the proportion of disease among those with both exposures that is attributable to their interaction (AP[AB]), and (3) the synergy index (S).

A RERI of one means no interaction or exactly additivity. A RERI of greater than one means positive interaction or more than additivity. A RERI of less than one means negative interaction or less than additivity. RERI ranges from zero to infinity.

An AP[AB] of zero means no interaction or exactly additivity. An AP[AB] greater than zero means positive interaction or more than additivity. An AP[AB] of less than zero means negative interaction or less than additivity. AP[AB] ranges from -1 to +1.

The synergy index is the ratio of the combined effects and the individual effects. An S of one means no interaction or exactly additivity. An S of greater than one means positive interaction or more than additivity. An S of less than one means negative interaction or less than additivity. S ranges from zero to infinity.

In the absence of interaction AP[AB] = 0 and RERI and S = 1.

Skrondal (2003) advocates for use of the synergy index as a summary measure of additive interaction, showing that when regression models adjust for the effect of confounding variables (as in the majority of cases) RERI and AP may be biased, while S remains unbiased.

This function uses the delta method to calculate the confidence intervals for each of the interaction measures, as described by Hosmer and Lemeshow (1992). An error will be returned if the point estimate of the synergy index is less than one. In this situation a warning is issued advising the user to re-parameterise their model as a linear odds model. See Skrondal (2003) for details.

A measure of multiplicative interaction is RR11 / (RR10 * RR01). If RR11 / (RR10 * RR01) equals one means no multiplicative interaction. If RR11 / (RR10 * RR01) is greater than one multiplicative interaction is said to be positive. If RR11 / (RR10 * RR01) is less than one multiplicative interaction is said to be negative.
epi.interaction

Value

A data frame listing:

- **est**: the point estimate of the requested additive interaction measure.
- **lower**: the lower bound of the confidence interval of the requested additive interaction measure.
- **upper**: the upper bound of the confidence interval of the requested additive interaction measure.

References


Examples

```r
## EXAMPLE 1:
## Data from Rothman and Keller (1972) evaluating the effect of joint exposure
to alcohol and tabacco on risk of cancer of the mouth and pharynx (cited in
## Hosmer and Lemeshow, 1992):

can <- c(rep(1, times = 231), rep(0, times = 178), rep(1, times = 11),
        rep(0, times = 38))
smk <- c(rep(1, times = 225), rep(0, times = 6), rep(1, times = 166),
        rep(0, times = 12), rep(1, times = 8), rep(0, times = 3), rep(1, times = 18),
        rep(0, times = 20))
alc <- c(rep(1, times = 409), rep(0, times = 49))
dat.df01 <- data.frame(alc, smk, can)

## Table 2 of Hosmer and Lemeshow (1992):
dat.glm01 <- glm(can ~ alc + smk + alc:smk, family = binomial, data = dat.df01)
summary(dat.glm01)

## What is the measure of effect modification on the additive scale?
epi.interaction(model = dat.glm01, param = "product", coef = c(2,3,4),
                 type = "RERI", conf.level = 0.95)
```
## Measure of interaction on the additive scale: RERI 3.73
## (95% CI -1.84 to 9.32), page 453 of Hosmer and Lemeshow (1992).

## What is the measure of effect modification on the multiplicative scale?
## for details.

```r
beta1 <- as.numeric(dat.glm01$coefficients[2])
beta2 <- as.numeric(dat.glm01$coefficients[3])
beta3 <- as.numeric(dat.glm01$coefficients[4])
exp(beta3) / (exp(beta1) * exp(beta2))
```

## Measure of interaction on the multiplicative scale: 0.093.

## EXAMPLE 2:
## Rothman defines an alternative coding scheme to be employed for
## parameterising an interaction term. Using this approach, instead of using
## two risk factors and one product term to represent the interaction (as
## above) the risk factors are combined into one variable comprised of
## (in this case) four levels:

```r
# a.neg b.neg: 0 0 0
# a.pos b.neg: 1 0 0
# a.neg b.pos: 0 1 0
# a.pos b.pos: 0 0 1

dat.df01$d <- rep(NA, times = nrow(dat.df01))
dat.df01$d[dat.df01$alc == 0 & dat.df01$smk == 0] <- 0
dat.df01$d[dat.df01$alc == 1 & dat.df01$smk == 0] <- 1
dat.df01$d[dat.df01$alc == 0 & dat.df01$smk == 1] <- 2
dat.df01$d[dat.df01$alc == 1 & dat.df01$smk == 1] <- 3
dat.df01$d <- factor(dat.df01$d)
```

## Table 3 of Hosmer and Lemeshow (1992):
##
## ```r
dat.glm02 <- glm(can ~ d, family = binomial, data = dat.df01)
summary(dat.glm02)
```

## What is the measure of effect modification on the additive scale?

```r
epi.interaction(model = dat.glm02, param = "dummy", coef = c(2,3,4),
    type = "RERI", conf.level = 0.95)
```

## Measure of interaction on the additive scale: RERI 3.73
## (95% CI -1.84 to 9.32), page 455 of Hosmer and Lemeshow (1992).
Description

Computes individual study odds or risk ratios for binary outcome data. Computes the summary odds or risk ratio using the inverse variance method. Performs a test of heterogeneity among trials. Performs a test for the overall difference between groups (that is, after pooling the studies, do treated groups differ significantly from controls?).

Usage

epi.iv(ev.trt, n.trt, ev.ctrl, n.ctrl, names, method = "odds.ratio",
alternative = c("two.sided", "less", "greater"), conf.level = 0.95)

Arguments

- ev.trt: observed number of events in the treatment group.
- n.trt: number in the treatment group.
- ev.ctrl: observed number of events in the control group.
- n.ctrl: number in the control group.
- names: character string identifying each trial.
- method: a character string indicating the method to be used. Options are odds.ratio or risk.ratio.
- alternative: a character string specifying the alternative hypothesis, must be one of two.sided, greater or less.
- conf.level: magnitude of the returned confidence interval. Must be a single number between 0 and 1.

Details

Using this method, the inverse variance weights are used to compute the pooled odds ratios and risk ratios. The inverse variance weights should be used to indicate the weight each trial contributes to the meta-analysis.

alternative = "greater" tests the hypothesis that the inverse variance summary measure of association is greater than 1.

Value

A list containing:

- OR: the odds ratio for each trial and the lower and upper bounds of the confidence interval of the odds ratio for each trial.
- RR: the risk ratio for each trial and the lower and upper bounds of the confidence interval of the risk ratio for each trial.
- OR.summary: the inverse variance summary odds ratio and the lower and upper bounds of the confidence interval of the inverse variance summary odds ratio.
- RR.summary: the inverse variance summary risk ratio and the lower and upper bounds of the confidence interval of the inverse variance summary risk ratio.
- weights: the raw and inverse variance weights assigned to each trial.
heterogeneity  a vector containing \( Q \) the heterogeneity test statistic, \( df \) the degrees of freedom and its associated P-value.

\( H_{sq} \)  the relative excess of the heterogeneity test statistic \( Q \) over the degrees of freedom \( df \).

\( I_{sq} \)  the percentage of total variation in study estimates that is due to heterogeneity rather than chance.

effect  a vector containing \( z \) the test statistic for overall treatment effect and its associated P-value.

Note

The inverse variance method performs poorly when data are sparse, both in terms of event rates being low and trials being small. The Mantel-Haenszel method (epi.mh) is more robust when data are sparse.

Using this method, the inverse variance weights are used to compute the pooled odds ratios and risk ratios.

The function checks each strata for cells with zero frequencies. If a zero frequency is found in any cell, 0.5 is added to all cells within the strata.

References


See Also
epi.dsl, epi.mh, epi.smd

Examples

```r
## EXAMPLE 1:
data(epi.epidural)
epi.iv(ev.trt = epi.epidural$ev.trt, n.trt = epi.epidural$n.trt,
      ev.ctrl = epi.epidural$ev.ctrl, n.ctrl = epi.epidural$n.ctrl,
      names = as.character(epi.epidural$trial), method = "odds.ratio",
      alternative = "two.sided", conf.level = 0.95)
```
epi.kappa

Kappa statistic

Description
Computes the kappa statistic and its confidence interval.

Usage
epi.kappa(dat, method = "fleiss", alternative = c("two.sided", "less", "greater"), conf.level = 0.95)

Arguments
dat an object of class matrix comprised of n rows and n columns listing the individual cell frequencies.
method a character string indicating the method to use. Options are fleiss, watson or altman.
alternative a character string specifying the alternative hypothesis, must be one of two.sided, greater or less.
conf.level magnitude of the returned confidence interval. Must be a single number between 0 and 1.

Details
Kappa is a measure of agreement beyond the level of agreement expected by chance alone. The observed agreement is the proportion of samples for which both methods (or observers) agree. The bias and prevalence adjusted kappa (Byrt et al. 1993) provides a measure of observed agreement, an index of the bias between observers, and an index of the differences between the overall proportion of 'yes' and 'no' assessments. The bias and prevalence adjusted kappa are only returned if the number of rows and columns of argument dat equal 2.

Common interpretations for the kappa statistic are as follows: < 0.2 slight agreement, 0.2 - 0.4 fair agreement, 0.4 - 0.6 moderate agreement, 0.6 - 0.8 substantial agreement, > 0.8 almost perfect agreement.

The argument alternative = "greater" tests the hypothesis that kappa is greater than 0.

Value
Where the number of rows and columns of argument dat is greater than 2 a list containing the following:

prop.agree a data frame with obs the observed proportion of agreement and exp the expected proportion of agreement.
pabak a data frame with the prevalence and bias corrected kappa statistic and the lower and upper bounds of the confidence interval for the prevalence and bias corrected kappa statistic.
kappa  a data frame with the kappa statistic, the standard error of the kappa statistic and the lower and upper bounds of the confidence interval for the kappa statistic.

z  a data frame containing the z test statistic for kappa and its associated P-value.

Where the number of rows and columns of argument dat is equal to 2 a list containing the following:

prop.agree  a data frame with obs the observed proportion of agreement and exp the expected proportion of agreement.

pindex  a data frame with the prevalence index, the standard error of the prevalence index and the lower and upper bounds of the confidence interval for the prevalence index.

bindex  a data frame with the bias index, the standard error of the bias index and the lower and upper bounds of the confidence interval for the bias index.

pabak  a data frame with the prevalence and bias corrected kappa statistic and the lower and upper bounds of the confidence interval for the prevalence and bias corrected kappa statistic.

kappa  a data frame with the kappa statistic, the standard error of the kappa statistic and the lower and upper bounds of the confidence interval for the kappa statistic.

z  a data frame containing the z test statistic for kappa and its associated P-value.

mcnemar  a data frame containing the McNemar test statistic for kappa and its associated P-value.

The kappa coefficient is influenced by the prevalence of the condition being assessed. A prevalence effect exists when the proportion of agreements on the positive classification differs from that of the negative classification. If the prevalence index is high (that is, the prevalence of a positive rating is very high or very low) chance agreement is also high and the value of kappa is reduced accordingly. The effect of prevalence on kappa is greater for large values of kappa than for small values (Byrt et al. 1993). Using the notation above, the prevalence index is calculated as \((a/N) - (d/N)\). Confidence intervals for the prevalence index are based on methods used for a difference in two proportions. See Rothman (2002, p 135 equation 7-2) for details.

Bias is the extent to which raters disagree on the proportion of positive (or negative) cases. Bias affects interpretation of the kappa coefficient. When there is a large amount of bias, kappa is higher than when bias is low or absent. In contrast to prevalence, the effect of bias is greater when kappa is small than when it is large (Byrt et al. 1993). Using the notation above, the bias index is calculated
as \( \frac{(a + b)}{N} - \frac{(a + c)}{N} \). Confidence intervals for the bias index are based on methods used for a difference in two proportions. See Rothman (2002, p 135 equation 7-2) for details.

The McNemar test is used to test for the presence of bias. A statistically significant McNemar test (generally if \( P < 0.05 \)) shows that there is evidence of a systematic difference between the proportion of 'positive' responses from the two methods. If one method provides the 'true values' (i.e. it is regarded as the gold standard method) the absence of a systematic difference implies that there is no bias. However, a non-significant result indicates only that there is no evidence of a systematic effect. A systematic effect may be present, but the power of the test may be inadequate to determine its presence.

References


Examples

```r
## EXAMPLE 1:
## Kidney samples from 291 salmon were split with one half of the
## samples sent to each of two laboratories where an IFAT test
## was run on each sample. The following results were obtained:

## Lab 1 positive, lab 2 positive: 19
## Lab 1 positive, lab 2 negative: 10
## Lab 1 negative, lab 2 positive: 6
## Lab 1 negative, lab 2 negative: 256

dat.m01 <- matrix(c(19,10,6,256), nrow = 2, byrow = TRUE)
colnames(dat.m01) <- c("L1-pos","L1-neg")
rownames(dat.m01) <- c("L2-pos","L2-neg")
epi.kappa(dat.m01, method = "fleiss", alternative = "greater")
```
## The z test statistic is 11.53 (P < 0.01). We accept the alternative hypothesis that the kappa statistic is greater than zero.

## The proportion of agreement after chance has been excluded is 0.67 (95% CI 0.56 to 0.79). We conclude that, on the basis of this sample, that there is substantial agreement between the two laboratories.

## EXAMPLE 2 (from Watson and Petrie 2010, page 1170):
Silva et al. (2007) compared an early pregnancy enzyme-linked immunosorbent assay test for pregnancy associated glycoprotein on blood samples collected from lactating dairy cows at day 27 after artificial insemination with transrectal ultrasound (US) diagnosis of pregnancy at the same stage.
The results were as follows:

ELISA positive, US positive: 596
ELISA positive, US negative: 61
ELISA negative, US positive: 29
ELISA negative, US negative: 987

data.m02 <- matrix(c(596, 61, 29, 987), nrow = 2, byrow = TRUE)
colnames(data.m02) <- c("US-pos", "US-neg")
rownames(data.m02) <- c("ELISA-pos", "ELISA-neg")

epi.kappa(data.m02, method = "watson", alternative = "greater", conf.level = 0.95)

## The proportion of agreements after chance has been excluded is 0.89 (95% CI 0.86 to 0.91). We conclude that there is substantial agreement between the two pregnancy diagnostic methods.

---

### epi.ltd

**Lactation to date and standard lactation milk yields**

#### Description

Calculate lactation to date and standard lactation (that is, 305 or 270 day) milk yields.

#### Usage

```
epi.ltd(dat, std = "305")
```

#### Arguments

- **dat**: an eight column data frame listing (in order) cow identifier, herd test identifier, lactation number, herd test days in milk, lactation length (NA if lactation incom-
plete), herd test milk yield (litres), herd test fat (percent), and herd test protein (percent).

std std = "305" returns 305-day milk volume, fat, and protein yield. std = "270" returns 270-day milk volume, fat, and protein yield.

Details

Lactation to date yields will only be calculated if there are four or more herd test events.

Value

A data frame with nine elements: ckey cow identifier, lact lactation number, llen lactation length, v1td milk volume (litres) to last herd test or dry off date (computed on the basis of lactation length, f1td fat yield (kilograms) to last herd test or dry off date (computed on the basis of lactation length, p1td protein yield (kilograms) to last herd test or dry off date (computed on the basis of lactation length, vstd 305-day or 270-day milk volume yield (litres), fstd 305-day or 270-day milk fat yield (kilograms), and pstd 305-day or 270-day milk protein yield (kilograms).

Author(s)

Nicolas Lopez-Villalobos (IVABS, Massey University, Palmerston North New Zealand) and Mark Stevenson (Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Australia).

References


Examples

## EXAMPLE 1:
## Generate some herd test data:
ckey <- rep(1, times = 12)
pkey <- 1:12
lact <- rep(1:2, each = 6)
dim <- c(25,68,105,145,200,240,30,65,90,130,190,220)
llen <- c(280,280,280,280,280,280,NA,NA,NA,NA,NA,NA)
vol <- c(18,30,25,68,105,145,200,240,30,65,90,130,190,220)
fat <- c(4.8,4.3,4.5,4.7,4.8,4.9,4.8,4.3,4.5,4.7,4.8,4.9)/100
pro <- c(3.7,3.5,3.6,3.7,3.8,3.9,3.7,3.5,3.6,3.7,3.8,3.9)/100
dat.df01 <- data.frame(ckey, pkey, lact, dim, llen, vol, fat, pro)

## Lactation to date and 305-day milk, fat, and protein yield:
epi.ltd(dat.df01, std = "305")

## Lactation to date and 270-day milk, fat, and protein yield:
epi.ltd(dat.df01, std = "270")
Description

Computes individual study odds or risk ratios for binary outcome data. Computes the summary odds or risk ratio using the Mantel-Haenszel method. Performs a test of heterogeneity among trials. Performs a test for the overall difference between groups (that is, after pooling the studies, do treated groups differ significantly from controls?).

Usage

```r
epi.mh(ev.trt, n.trt, ev.ctrl, n.ctrl, names, method = "odds.ratio", alternative = c("two.sided", "less", "greater"), conf.level = 0.95)
```

Arguments

- `ev.trt`: observed number of events in the treatment group.
- `n.trt`: number in the treatment group.
- `ev.ctrl`: observed number of events in the control group.
- `n.ctrl`: number in the control group.
- `names`: character string identifying each trial.
- `method`: a character string indicating the method to be used. Options are `odds.ratio` or `risk.ratio`.
- `alternative`: a character string specifying the alternative hypothesis, must be one of `two.sided`, `greater` or `less`.
- `conf.level`: magnitude of the returned confidence interval. Must be a single number between 0 and 1.

Details

`alternative = "greater"` tests the hypothesis that the Mantel-Haenszel summary measure of association is greater than 1.

Value

A list containing the following:

- `OR`: the odds ratio for each trial and the lower and upper bounds of the confidence interval of the odds ratio for each trial.
- `RR`: the risk ratio for each trial and the lower and upper bounds of the confidence interval of the risk ratio for each trial.
- `OR.summary`: the Mantel-Haenszel summary odds ratio and the lower and upper bounds of the confidence interval of the Mantel-Haenszel summary odds ratio.
RR.summary = the Mantel-Haenszel summary risk ratio and the lower and upper bounds of the confidence interval of the Mantel-Haenszel summary risk ratio.
weights = the raw and inverse variance weights assigned to each trial.
heterogeneity = a vector containing Q the heterogeneity test statistic, df the degrees of freedom and its associated P-value.
Hsq = the relative excess of the heterogeneity test statistic Q over the degrees of freedom df.
Isq = the percentage of total variation in study estimates that is due to heterogeneity rather than chance.
effect = a vector containing z the test statistic for overall treatment effect and its associated P-value.

Note
Using this method, the pooled odds and risk ratios are computed using the raw individual study weights. The methodology for computing the Mantel-Haenszel summary odds ratio follows the approach described in Deeks, Altman and Bradburn MJ (2001, pp 291 - 299).
The function checks each strata for cells with zero frequencies. If a zero frequency is found in any cell, 0.5 is added to all cells within the strata.

References

See Also
epi.dsl, epi.iv, epi.smd

Examples
## EXAMPLE 1:
data(epi.epidural)
epi.mh(ev.trt = epi.epidural$ev.trt, n.trt = epi.epidural$n.trt,
ev.ctrl = epi.epidural$ev.ctrl, n.ctrl = epi.epidural$n.ctrl,
names = as.character(epi.epidural$trial), method = "odds.ratio",
alternative = "two.sided", conf.level = 0.95)
epi.nomogram

Post-test probability of disease given sensitivity and specificity of a test

Description

Compute the post-test probability of disease given sensitivity and specificity of a test.

Usage

epi.nomogram(se, sp, lr, pre.pos, verbose = FALSE)

Arguments

se    test sensitivity (0 - 1).
sp    test specificity (0 - 1).
lr    a vector of length 2 listing the positive and negative likelihood ratio (respectively) of the test. Ignored if se and sp are not null.
pre.pos    the pre-test probability of the outcome.
verbose    logical, indicating whether detailed or summary results are to be returned.

Value

A list containing the following:

lr    a data frame listing the likelihood ratio of a positive and negative test.
prior    a data frame listing the pre-test probability of being outcome (i.e. disease) positive, as entered by the user.
post    a data frame listing: opos.tpos the post-test probability of being outcome (i.e. disease) positive given a positive test result and opos.tneg the post-test probability of being outcome (i.e. disease) positive given a negative test result.

References


Examples

## EXAMPLE 1:
## You are presented with a dog with lethargy, exercise intolerance, weight gain and bilaterally symmetric truncal alopecia. You are suspicious of hypothyroidism and take a blood sample to measure basal serum thyroxine (T4).
## You believe that around 5% of dogs presented to your clinic with a signalment of general debility have hypothyroidism. The serum T4 has a sensitivity of 0.89 and specificity of 0.85 for diagnosing hypothyroidism in the dog. The laboratory reports a serum T4 concentration of 22.0 nmol/L (reference range 19.0 to 58.0 nmol/L). What is the post-test probability that this dog is hypothyroid?

```r
epi.nomogram(se = 0.89, sp = 0.85, lr = NA, pre.pos = 0.05, verbose = FALSE)
```

## If the test is positive the post-test probability that this dog is hypothyroid is 0.24. If the test is negative the post-test probability that this dog is hypothyroid is 0.0068.

## EXAMPLE 2:
## A dog is presented to you with severe pruritis. You suspect sarcoptic mange and decide to take a skin scraping (LR+ 9000; LR- 0.1). The scrape returns a negative result (no mites are seen). What is the post-test probability that your patient has sarcoptic mange? You recall that you diagnose around 3 cases of sarcoptic mange per year in a clinic that sees approximately 2 -- 3 dogs per week presented with pruritic skin disease.

## Calculate the pre-test probability of sarcoptes:
```r
pre.pos <- 3 / (3 * 52)
```
## The pre-test probability that this dog is sarcoptes positive is 0.019.

```r
epi.nomogram(se = NA, sp = NA, lr = c(9000, 0.1), pre.pos = pre.pos, verbose = FALSE)
```

## If the skin scraping is negative the post-test probability that this dog has sarcoptic mange is 0.002.

---

### epi.occc

**Overall concordance correlation coefficient (OCCC)**

**Description**

Overall concordance correlation coefficient (OCCC) for agreement on a continuous measure based on Lin (1989, 2000) and Barnhart et al. (2002).

**Usage**

```r
epi.occc(dat, na.rm = FALSE, pairs = FALSE)
```

## S3 method for class 'epi.occc'
```r
print(x, ...)
```

## S3 method for class 'epi.occc'
```r
summary(object, ...)
```
Arguments

dat a matrix, or a matrix like object. Rows correspond to cases/observations, columns
    corresponds to raters/variables.
na.rm logical. Should missing values (including NaN) be removed?
pairs logical. Should the return object contain pairwise statistics? See Details.
x, object an object of class epi.occc.
... further arguments passed to print methods.

Details

The index proposed by Barnhart et al. (2002) is the same as the index suggested by Lin (1989) in
the section of future studies with a correction of a typographical error in Lin (2000).

Value

An object of class epi.occc with the following list elements (notation follows Barnhart et al. 2002):

• occc: the value of the overall concordance correlation coefficient ($\rho_{oc}$),
• oprec: overall precision ($\rho$),
• oaccu: overall accuracy ($\chi^a$),
• pairs: a list with following elements (only if pairs = TRUE, otherwise NULL; column indices
    for the pairs (j,k) follow lower-triangle column-major rule based on a ncol(x) times ncol(x)
    matrix),
    – ccc: pairwise CCC values ($\rho_{jk}$),
    – prec: pairwise precision values ($\rho_{jk}$),
    – accu: pairwise accuracy values ($\chi^a_{jk}$),
    – ksi: pairwise weights ($\xi_{jk}$),
    – scale: pairwise scale values ($v_{jk}$),
    – location: pairwise location values ($u_{jk}$),
• data.name: name of the input data dat.

Author(s)

Peter Solymos, solymos@ualberta.ca.

References

Barnhart H X, Haber M, Song J (2002). Overall concordance correlation coefficient for evaluating
- 268.

See Also

epi.ccc
Examples

```r
## EXAMPLE 1:
## Generate some rating data:
set.seed(1234)
p <- runif(n = 10, min = 0, max = 1)
x <- replicate(n = 5, expr = rbinom(n = 10, size = 4, prob = p) + 1)

rval.occc01 <- epi.occc(dat = x, pairs = TRUE)
print(rval.occc01); summary(rval.occc01)
```

---

epi.offset

Create offset vector

Description

Creates an offset vector based on a list.

Usage

epi.offset(id.names)

Arguments

- `id.names`: a list identifying the [location] of each case. This must be a factor.

Details

This function is useful for supplying spatial data to WinBUGS.

Value

A vector of length \((1 + \text{length of } \text{id})\). The first element of the offset vector is 1, corresponding to the position at which data for the first factor appears in id. The second element of the offset vector corresponds to the position at which the second factor appears in id and so on. The last element of the offset vector corresponds to the length of the id list.

References


Examples

```r
## EXAMPLE 1:
dat.v01 <- c(1,1,1,2,2,2,2,3,3,3)
dat.v01 <- as.factor(dat.v01)
dat.ofs01 <- epi.offset(dat.v01)
dat.ofs01
## [1] 1 4 8 10
```

epi.pooled  

Estimate herd test characteristics when pooled sampling is used

Description

We may wish to designate a group of individuals (e.g. a herd) as being either diseased or non-diseased on the basis of pooled samples. This function estimates sensitivity and specificity of this testing regime at the group (or herd) level.

Usage

```r
epi.pooled(se, sp, P, m, r)
```

Arguments

- `se`: a vector of length one defining the sensitivity of the individual test used.
- `sp`: a vector of length one defining the specificity of the individual test used.
- `P`: scalar, defining the estimated true prevalence.
- `m`: scalar, defining the number of individual samples to make up a pooled sample.
- `r`: scalar, defining the number of pooled samples per group (or herd).

Value

A list containing the following:

- `HAPneg`: the apparent prevalence in a disease negative herd.
- `HSe`: the estimated group (herd) level sensitivity.
- `HSp`: the estimated group (herd) level specificity.

References


Examples

```r
## EXAMPLE 1:
## We want to test dairy herds for Johne's disease using faecal culture
## which has a sensitivity and specificity of 0.647 and 0.981, respectively.
## Suppose we pool faecal samples from five cows together and collect six
## pooled samples per herd. What is the herd level sensitivity and specificity
## based on this approach (assuming homogenous mixing)?

epi.pooled(se = 0.647, sp = 0.981, P = 0.12, m = 5, r = 6)

## Herd level sensitivity is 0.927, herd level specificity is 0.562.
## Sensitivity at the herd level is increased using the pooled sampling
## approach. Herd level specificity is decreased.
```

---

**epi.popsize**

*Estimate population size on the basis of capture-recapture sampling*

### Description

Estimates population size on the basis of capture-recapture sampling.

### Usage

```r
epi.popsize(T1, T2, T12, conf.level = 0.95, verbose = FALSE)
```

### Arguments

- **T1**: an integer representing the number of individuals tested in the first round.
- **T2**: an integer representing the number of individuals tested in the second round.
- **T12**: an integer representing the number of individuals tested in both the first and second round.
- **conf.level**: magnitude of the returned confidence interval. Must be a single number between 0 and 1.
- **verbose**: logical indicating whether detailed or summary results are to be returned.

### Value

Returns the estimated population size and an estimate of the numbers of individuals that remain untested.

### References

Examples

## EXAMPLE 1:
## In a field survey 400 feral pigs are captured, marked and then released.
## On a second occasion 40 of the original capture are found when another 400
## pigs are captured. Estimate the size of this feral pig population. Estimate
## the number of feral pigs that have not been tested.

epi.popsizet(T1 = 400, T2 = 400, T12 = 40, conf.level = 0.95, verbose = FALSE)

## Estimated population size: 4000 (95% CI 3125 to 5557)
## Estimated number of untested pigs: 3240 (95% CI 2365 to 4797)

### Description

Compute partial rank correlation coefficients.

### Usage

epi.prcc(dat, sided.test = 2, conf.level = 0.95)

### Arguments

dat: a data frame comprised of $K + 1$ columns and $N$ rows, where $K$ represents the number of model parameters being evaluated and $N$ represents the number of replications of the model. The last column of the data frame (i.e. column $K + 1$) provides the model output.

sided.test: use a one- or two-sided test? Use a two-sided test if you wish to evaluate whether or not the partial rank correlation coefficient is greater than or less than zero. Use a one-sided test to evaluate whether or not the partial rank correlation coefficient is greater than zero.

conf.level: magnitude of the returned confidence intervals. Must be a single number between 0 and 1.

### Details

Calculation of the PRCC enables the determination of the statistical relationships between each input parameter and the outcome variable while keeping all of the other input parameters constant at their expected value (Conover, 1980). This procedure enables the independent effects of each parameter to be determined, even when the parameters are correlated. A PRCC indicates the degree of monotonicity between a specific input variable and an outcome; therefore only outcome variables that are monotonically related to the input parameters should be chosen for this analysis (Conover, 1980; Iman and Conover 1980). Monotonicity can be assessed by examining scatterplots where each input variable is plotted as a function of the outcome variable. The sign of the PRCC indicates
the qualitative relationship between each input variable and the outcome variable. The magnitude of the PRCC indicates the importance of the uncertainty in the input variable in contributing to the imprecision in predicting the value of the outcome variable. The relative importance of the input variables can be directly evaluated by comparing the values of the PRCC.

If the number of parameters \( k \) is greater than the number of model replications \( N \) an error will be returned.

**Value**

A data frame with three elements: `est` the point estimate of the partial rank correlation coefficient between each input parameter and the outcome, `lower` the lower bound of the confidence interval of the partial rank correlation coefficient, `upper` the upper bound of the confidence interval of the partial rank correlation coefficient, `test.statistic` the test statistic used to determine the significance of non-zero values of the partial rank correlation coefficient, and `p.value` the associated P-value.

**Author(s)**

Jonathon Marshall, J.C.Marshall@massey.ac.nz.

**References**


**Examples**

```r
## EXAMPLE 1:
## Create a matrix of simulation results:
x1 <- rnorm(n = 10, mean = 120, sd = 130)
x2 <- rnorm(n = 10, mean = 80, sd = 5)
x3 <- rnorm(n = 10, mean = 40, sd = 20)
y <- 2 + (0.5 * x1) - (1.7 * x2) + (0.2 * x3)
dat.df01 <- data.frame(x1 = x1, x2 = x2, x3 = x3, y = y)
epi.prcc(dat.df01, sided.test = 2, conf.level = 0.95)
```
**epi.prev**

**Estimate true prevalence**

**Description**

Compute the true prevalence of a disease in a population on the basis of an imperfect test.

**Usage**

```r
epi.prev(pos, tested, se, sp, method = "wilson", units = 100, conf.level = 0.95)
```

**Arguments**

- `pos`: a vector listing the count of positive test results for each population.
- `tested`: a vector listing the count of subjects tested for each population.
- `se`: test sensitivity (0 - 1). `se` can either be a single number or a vector of the same length as `pos`. See the examples, below, for details.
- `sp`: test specificity (0 - 1). `sp` can either be a single number or a vector of the same length as `pos`. See the examples, below, for details.
- `method`: a character string indicating the confidence interval calculation method to use. Options are "c-p" (Clopper-Pearson), "sterne" (Sterne), "blaker" (Blaker) and "wilson" (Wilson).
- `units`: multiplier for the prevalence estimates.
- `conf.level`: magnitude of the returned confidence interval. Must be a single number between 0 and 1.

**Details**

Appropriate confidence intervals for the adjusted prevalence estimate are provided, accounting for the change in variance that arises from imperfect test sensitivity and specificity (see Reiczigel et al 2010 for details).

The Clopper-Pearson method is known to be too conservative for two-sided intervals (Blaker 2000, Agresti and Coull 1998). Blaker’s and Sterne’s methods (Blaker 2000, Sterne 1954) provide smaller exact two-sided confidence interval estimates.

**Value**

A list containing the following:

- `ap`: the point estimate of apparent prevalence and the lower and upper bounds of the confidence interval around the apparent prevalence estimate.
- `tp`: the point estimate of the true prevalence and the lower and upper bounds of the confidence interval around the true prevalence estimate.
Note

This function uses apparent prevalence, test sensitivity and test specificity to estimate true prevalence (after Rogan and Gladen, 1978). Confidence intervals for the apparent and true prevalence estimates are based on code provided by Reiczigel et al. (2010).

If apparent prevalence is less than \((1 - \text{diagnostic test specificity})\) the Rogan Gladen estimate of true prevalence will be less than zero (Speybroeck et al. 2012). If the apparent prevalence is greater than the diagnostic test sensitivity the Rogan Gladen estimate of true prevalence will be greater than one.

When \(\text{AP} < (1 - \text{Sp})\) the function issues a warning to alert the user that the estimate of true prevalence is invalid. A similar warning is issued when \(\text{AP} > \text{Se}\). In either situation a Bayesian approach for estimation of true prevalence is recommended. See Messam et al. (2008) for a concise introduction to this topic.

References


Examples

```r
## EXAMPLE 1:
## A simple random sample of 150 cows from a herd of 2560 is taken.
## Each cow is given a screening test for brucellosis which has a
## sensitivity of 96% and a specificity of 89%. Of the 150 cows tested
## 45 were positive to the screening test. What is the estimated prevalence
## of brucellosis in this herd (and its 95% confidence interval)?
epi.prev(pos = 45, tested = 150, se = 0.96, sp = 0.89, method = "blaker",
         units = 100, conf.level = 0.95)
```
## The estimated true prevalence of brucellosis in this herd is 22 (95% 14 to 32) cases per 100 cows at risk.

# EXAMPLE 2:
### Moujaber et al. (2008) analysed the seroepidemiology of Helicobacter pylori infection in Australia. They reported seroprevalence rates together with 95% confidence intervals by age group using the Clopper-Pearson exact method (Clopper and Pearson, 1934). The ELISA test they applied had 96.4% sensitivity and 92.7% specificity. A total of 151 subjects 1 -- 4 years of age were tested. Of this group 6 were positive. What is the estimated true prevalence of Helicobacter pylori in this age group?

```r
epi.prev(pos = 6, tested = 151, se = 0.964, sp = 0.927, method = "c-p", units = 100, conf.level = 0.95)
```

## The estimated true prevalence of Helicobacter pylori in 1 -- 4 year olds is -4 (95% CI -6 to 1) cases per 100. The function issues a warning to alert the user that estimate of true prevalence invalid.

## EXAMPLE 3:
### Three dairy herds are tested for tuberculosis. On each herd a different test regime is used (each with a different diagnostic test sensitivity and specificity). The number of animals tested in each herd were 210, 189 and 124, respectively. The number of test-positives in each herd were 8, 12 and 7. Test sensitivities were 0.60, 0.65 and 0.70 (respectively). Test specificities were 0.90, 0.95 and 0.99. What is the estimated true prevalence of tuberculosis in the three herds?

```r
rval.prev03 <- epi.prev(pos = c(80,100,50), tested = c(210,189,124), se = c(0.60,0.65,0.70), sp = c(0.90,0.95,0.99), method = "blaker", units = 100, conf.level = 0.95)
round(rval.prev03$tp, digits = 0)
```

### True prevalence estimates for each herd:
**Herd 1:** 56 (95% CI 43 to 70) cases per 100 cows.  
**Herd 2:** 80 (95% CI 68 to 92) cases per 100 cows.  
**Herd 3:** 57 (95% CI 45 to 70) cases per 100 cows.

### epi.psi

#### Proportional similarity index

**Description**

Compute proportional similarity index.
Usage

epi.psi(dat, itno = 99, conf.level = 0.95)

Arguments

dat  a data frame providing details of the distributions to be compared (in columns). The first column (either a character or factor) lists the levels of each distribution. Additional columns list the number of events for each factor level for each distribution to be compared.

itno  scalar, numeric defining the number of bootstrap simulations to be run to generate a confidence interval around the proportional similarity index estimate.

conf.level  scalar, numeric defining the magnitude of the returned confidence interval for each proportional similarity index estimate.

Details

The proportional similarity or Czekanowski index is an objective and simple measure of the area of intersection between two non-parametric frequency distributions (Feinsinger et al. 1981). PIS values range from 1 for identical frequency distributions to 0 for distributions with no common types. Bootstrap confidence intervals for this measure are estimated based on the approach developed by Garrett et al. (2007).

Value

A five column data frame listing: v1 the name of the reference column, v2 the name of the comparison column, est the estimated proportional similarity index, lower the lower bound of the estimated proportional similarity index, and upper the upper bound of the estimated proportional similarity index.

References


## EXAMPLE 1:

A cross-sectional study of Australian thoroughbred race horses was carried out. The sampling frame for this study comprised all horses registered with Racing Australia in 2017 -- 2018. A random sample of horses was selected from the sampling frame and the owners of each horse invited to take part in the study. Counts of source population horses and study population horses are provided below. How well did the geographic distribution of study population horses match the source population?

```r
state <- c("NSW", "VIC", "QLD", "WA", "SA", "TAS", "NT", "Abroad")
srcp <- c(11372, 10722, 7371, 4200, 2445, 1029, 510, 101)
stup <- c(622, 603, 259, 105, 102, 37, 22, 0)
dat.df01 <- data.frame(state, srcp, stup)
epi.psi(dat.df01, itno = 99, conf.level = 0.95)
```

The proportional similarity index for these data was 0.88 (95% CI 0.86 to 0.90). We conclude that the distribution of sampled horses by state was consistent with the distribution of the source population by state.

## Not run:

Compare the relative frequencies of the source and study populations by state graphically:

```r
library(ggplot2)

dat.df01$psrcp <- dat.df01$srcp / sum(dat.df01$srcp)
dat.df01$pstup <- dat.df01$stup / sum(dat.df01$stup)
dat.df01 <- dat.df01[sort.list(dat.df01$psrcp),]
dat.df01$state <- factor(dat.df01$state, levels = dat.df01$state)

gdat.df01 <- data.frame(state = rep(dat.df01$state, times = 2),
                       pop = c(rep("Source", times = nrow(dat.df01)),
                               rep("Study", times = nrow(dat.df01))),
                       pfreq = c(dat.df01$psrcp, dat.df01$pstup))
gdat.df01$state <- factor(gdat.df01$state, levels = dat.df01$state)

ggplot(data = gdat.df01, aes(x = state, y = pfreq)) +
  geom_bar(stat = "identity", position = position_dodge(), color = "grey") +
  scale_x_discrete(name = "State") +
  scale_y_continuous(limits = c(0, 0.50), name = "Proportion")
```

## End(Not run)
Description

Writes data from an R list to a text file in WinBUGS-compatible format.

Usage

epi.RtoBUGS(datalist, towhere)

Arguments

datalist    a list (normally, with named elements) which may include scalars, vectors, ma-
            trices, arrays of any number of dimensions, and data frames.
towhere    a character string identifying where the file is to be written.

Details

The function doesn’t check to ensure that only numbers are being produced. In particular, factor
labels in a dataframe will be output to the file, which normally won’t be desired.

Author(s)

Terry Elrod (terry.elrod@ualberta.ca), Kenneth Rice.

References

Best, NG. WinBUGS 1.3.1 Short Course, Brisbane, November 2000.

---


Description

This data set provides counts of lip cancer diagnoses made in Scottish districts from 1975 to 1980.
In addition to district-level counts of disease events and estimates of the size of the population at
risk, the data set contains (for each district) an estimate of the percentage of the population involved
in outdoor industry (agriculture, fishing, and forestry). It is known that exposure to sunlight is a
risk factor for cancer of the lip and high counts are to be expected in districts where there is a high
proportion of the workforce involved in outdoor industry.

Usage

data(epi.SClip)
Format

A data frame with 56 observations on the following 6 variables.

- **gridcode**: alternative district identifier.
- **id**: numeric district identifier (1 to 56).
- **district**: district name.
- **cases**: number of lip cancer cases diagnosed 1975 - 1980.
- **population**: total person years at risk 1975 - 1980.
- **prop.ag**: percent of the population engaged in outdoor industry.

Source

This data set has been analysed by a number of authors including Clayton and Kaldor (1987), Conlon and Louis (1999), Stern and Cressie (1999), and Carlin and Louis (2000, p 270).

References


epi.smd

Fixed-effects meta-analysis of continuous outcomes using the standardised mean difference method

Description

Computes the standardised mean difference and confidence intervals of the standardised mean difference for continuous outcome data.

Usage

```r
epi.smd(mean.trt, sd.trt, n.trt, mean.ctrl, sd.ctrl, n.ctrl,
        names, method = "cohens", conf.level = 0.95)
```
Arguments

- **mean.trt**: a vector, defining the mean outcome in the treatment group.
- **sd.trt**: a vector, defining the standard deviation of the outcome in the treatment group.
- **n.trt**: a vector, defining the number of subjects in the treatment group.
- **mean.ctrl**: a vector, defining the mean outcome in the control group.
- **sd.ctrl**: a vector, defining the standard deviation of the outcome in the control group.
- **n.ctrl**: a vector, defining the number of subjects in the control group.
- **names**: character string identifying each trial.
- **method**: a character string indicating the method to be used. Options are cohens or hedges and glass.
- **conf.level**: magnitude of the returned confidence interval. Must be a single number between 0 and 1.

Value

A list containing the following:

- **md**: standardised mean difference and its confidence interval computed for each trial.
- **md.invar**: the inverse variance (fixed effects) summary standardised mean difference.
- **md.dsl**: the DerSimonian and Laird (random effects) summary standardised mean difference.
- **heterogeneity**: a vector containing Q the heterogeneity test statistic, df the degrees of freedom and its associated P-value.

Note

The standardised mean difference method is used when trials assess the same outcome, but measure it in a variety of ways. For example: a set of trials might measure depression scores in psychiatric patients but use different methods to quantify depression. In this circumstance it is necessary to standardise the results of the trials to a uniform scale before they can be combined. The standardised mean difference method expresses the size of the treatment effect in each trial relative to the variability observed in that trial.

References


See Also

- epi.dsl, epi.iv, epi.mh
Examples

## EXAMPLE 1:

## A systematic review comparing assertive community treatment (ACT) for the
## severely mentally ill was compared to standard care. A systematic review
## comparing ACT to standard care found three trials that assessed mental
## status after 12 months. All three trials used a different scoring system,
## so standardisation is required before they can be compared.

```r
names <- c("Audini", "Morse", "Lehman")
mean.trt <- c(41.4, 0.95, -4.10)
mean.ctrl <- c(42.3, 0.89, -3.80)
sd.trt <- c(14, 0.76, 0.83)
sd.ctrl <- c(12.4, 0.65, 0.87)
n.trt <- c(30, 37, 67)
n.ctrl <- c(28, 35, 58)
epi.smd(mean.trt, sd.trt, n.trt, mean.ctrl, sd.ctrl, n.ctrl,
       names, method = "cohens", conf.level = 0.95)
```

### epi.smr

Confidence intervals and tests of significance of the standardised mortality [morbidity] ratio

Description

Computes confidence intervals and tests of significance of the standardised mortality [morbidity] ratio.

Usage

```r
epi.smr(obs, exp, method = "byar", conf.level = 0.95)
```

Arguments

- **obs**: scalar integer, defining the observed number of events.
- **exp**: scalar number, defining the expected number of events.
- **method**: character string, defining the method used. Options are `chi2`, `mid.p`, `fisher`, `byar`, `rothman.greenland`, `ury.wiggins` and `vandenbroucke`. See details, below.
- **conf.level**: magnitude of the returned confidence interval. Must be a single number between 0 and 1.

Details

This function calculates the standardised mortality [morbidity] ratio based on scalars defining the observed and expected number of [disease] events.
The hypothesis that the SMR equals one is tested using the Chi square test, the Mid-P exact test, the Fisher exact test and Byar’s approximation. Confidence intervals for the SMR are calculated using the Mid-P exact test, the Fisher exact test, Byar’s approximation, Rothman and Greenland’s method, Ury and Wiggin’s method and the Vandenbroucke method.

Exact confidence intervals and p-values should be used when the number of observed events is less than or equal to five. For greater numbers of observed events, the approximation methods (Byar’s, Rothman and Greenland, Ury and Wiggins and Vandenbroucke) should be used.

A two-sided test of significance is returned, using the test statistic appropriate for the method used.

Value

A data frame listing:

- **obs**: the observed number of events, as entered by the user.
- **exp**: the expected number of events, as entered by the user.
- **est**: the point estimate of the SMR.
- **lower**: the lower bound of the confidence interval of the SMR.
- **upper**: the upper bound of the confidence interval of the SMR.
- **test.statistic**: test statistic of the significance of the SMR.
- **p.value**: the probability that the null hypothesis (i.e. the number of observed events divided by the expected number of events equals 1) is true.

Note

Only 90%, 95% and 99% confidence limits are computed using the Ury and Wiggins method. If `conf.level` does not equal 0.90, 0.95 or 0.99 NAs are returned for the lower and upper bound of the SMR confidence interval.

Only 95% confidence limits are computed using Vandenbroucke’s method. If `conf.level` does not equal 0.95 NAs are returned for the lower and upper bound of the SMR confidence interval.

References

Examples

## EXAMPLE 1:
## The observed number of disease events in a province is 4; the expected
## number of disease events is 3.3. What is the standardised morbidity ratio
## and its 95% confidence interval? Test the hypothesis that the SMR equals
## one.

epi.smr(obs = 4, exp = 3.3, method = "mid.p", conf.level = 0.95)

## The standardised morbidity ratio is 1.2 (95% CI 0.38 to 2.9). We accept
## the null hypothesis and conclude that the SMR does not significantly
## differ from one (p = 0.657).

---

epi.ssc
c

Sample size, power or minimum detectable odds ratio for an un-
matched or matched case-control study

Description

Calculates the sample size, power or minimum detectable odds ratio for an unmatched or matched
case-control study.

Usage

epi.ssc(OR, p0, n, power, r = 1, rho.cc = 0, design = 1, sided.test = 2,
nfractional = FALSE, conf.level = 0.95, method = "unmatched", fleiss = FALSE)

Arguments

- **OR**
  scalar, the expected study odds ratio.

- **p0**
  scalar, the prevalence of exposure amongst the controls.

- **n**
  scalar, the total number of subjects in the study (i.e. the number of cases plus
  the number of controls).

- **power**
  scalar, the required study power.

- **r**
  scalar, the number in the control group divided by the number in the case group.

- **rho.cc**
  scalar, the correlation between case and control exposures for matched pairs.
  Ignored when method = "unmatched".

- **design**
  scalar, the design effect.

- **sided.test**
  use a one- or two-sided test? Use a two-sided test if you wish to evaluate whether
  or not the odds of exposure in cases is greater than or less than the odds of
  exposure in controls. Use a one-sided test to evaluate whether or not the odds of
  exposure in cases is greater than the odds of exposure in controls.

- **nfractional**
  logical, return fractional sample size.

- **conf.level**
  scalar, the level of confidence in the computed result.
method a character string defining the method to be used. Options are unmatched or matched.

fleiss logical, indicating whether or not the Fleiss correction should be applied. This argument is ignored when method = "matched".

Details

This function implements the methodology described by Dupont (1988). A detailed description of sample size calculations for case-control studies (with numerous worked examples, many of them reproduced below) is provided by Woodward (2005), pp. 381 to 426.

Value

A list containing the following:

- n.total the total number of subjects required to estimate the specified odds ratio at the desired level of confidence and power (i.e. the number of cases plus the number of controls).
- n.case the total number of case subjects required to estimate the specified odds ratio at the desired level of confidence and power.
- n.control the total number of control subjects required to estimate the specified odds ratio at the desired level of confidence and power.
- power the power of the study given the number of study subjects, the specified odds ratio and the desired level of confidence.
- OR the expected detectable odds ratio given the number of study subjects, the desired power and desired level of confidence.

Note

The power of a study is its ability to demonstrate the presence of an association, given that an association actually exists.

See the documentation for epi.sscohortc for an example using the design facility implemented in this function.

References


Examples

## EXAMPLE 1 (from Woodward 2005 p. 412):
## A case-control study of the relationship between smoking and CHD is
## planned. A sample of men with newly diagnosed CHD will be compared for
## smoking status with a sample of controls. Assuming an equal number of
## cases and controls, how many study subject are required to detect an
## odds ratio of 2.0 with 0.90 power using a two-sided 0.05 test? Previous
## surveys have shown that around 0.30 of males without CHD are smokers.

```r
epi.sscc(OR = 2.0, p0 = 0.30, n = NA, power = 0.90, r = 1, rho.cc = 0,
      design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95,
      method = "unmatched", fleiss = FALSE)
```

## A total of 376 men need to be sampled: 188 cases and 188 controls.

## EXAMPLE 2 (from Woodward 2005 p. 414):
## Suppose we wish to determine the power to detect an odds ratio of 2.0
## using a two-sided 0.05 test when 188 cases and 940 controls
## are available (that is, the ratio of controls to cases is 5:1). Assume
## the prevalence of smoking in males without CHD is 0.30.

```r
n <- 188 + 940
epi.sscc(OR = 2.0, p0 = 0.30, n = n, power = NA, r = 5, rho.cc = 0,
      design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95,
      method = "unmatched", fleiss = TRUE)
```

## The power of this study, with the given sample size allocation is 0.99.

## EXAMPLE 3:
## The following statement appeared in a study proposal to identify risk
## factors for campylobacteriosis in humans:

## 'We will prospectively recruit 300 culture-confirmed Campylobacter cases
## reported under the Public Health Act. We will then recruit one control per
## case from general practices of the enrolled cases, using frequency matching
## by age and sex. With exposure levels of 10% (thought to be realistic
## given past foodborne disease case control studies) this sample size
## will provide 80% power to detect an odds ratio of 2 at the 5% alpha
## level.'

## Confirm the statement that 300 case subjects will provide 80% power in
## this study.

```r
epi.sscc(OR = 2.0, p0 = 0.10, n = 600, power = NA, r = 1, rho.cc = 0.01,
      design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95,
      method = "matched", fleiss = TRUE)
```

## If the true odds ratio for Campylobacter in exposed subjects relative to
## unexposed subjects is 2.0 we will be able to reject the null hypothesis
## that this odds ratio equals 1 with probability (power) 0.826. The Type I
# error probability associated with this test of this null hypothesis is 0.05.

## EXAMPLE 4:
We wish to conduct a case-control study to assess whether bladder cancer may be associated with past exposure to cigarette smoking. Cases will be patients with bladder cancer and controls will be patients hospitalised for injury. It is assumed that 20% of controls will be smokers or past smokers, and we wish to detect an odds ratio of 2 with power 90%.
Three controls will be recruited for every case. How many subjects need to be enrolled in the study?

epi.sscc(OR = 2.0, p0 = 0.20, n = NA, power = 0.90, r = 3, rho.cc = 0, design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95, method = "unmatched", fleiss = FALSE)

## EXAMPLE 5:
Code to reproduce the isograph shown in Figure 2 in Dupont (1988):

r <- 1
p0 = seq(from = 0.05, to = 0.95, length = 50)
OR <- seq(from = 1.05, to = 6, length = 100)
dat.df05 <- expand.grid(p0 = p0, OR = OR)
dat.df05$n.total <- NA
for(i in 1:nrow(dat.df05)){
  dat.df05$n.total[i] <- epi.sscc(OR = dat.df05$OR[i], p0 = dat.df05$p0[i], n = NA, power = 0.80, r = 1, rho.cc = 0, design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95, method = "unmatched", fleiss = FALSE)$n.total
}
grid.n <- matrix(dat.df05$n.total, nrow = length(p0))
breaks <- c(22:30, 32, 34, 36, 40, 45, 50, 55, 60, 70, 80, 100, 125, 150, 175, 200, 300, 500, 1000)
par(mar = c(5,5,0,5), bty = "n")
contour(x = p0, y = OR, z = log10(grid.n), add = FALSE, levels = log10(breaks),
labels = breaks, xlim = c(0,1), ylim = c(1,6), las = 1, method = "flattest",
xlab = 'Proportion of controls exposed', ylab = "Minimum OR to detect")

## Not run:
## The same plot using ggplot2:
library(ggplot2); library(directlabels)
p <- ggplot(data = dat.df05, aes(x = p0, y = OR, z = n.total)) +
  geom_contour(aes(colour = ..level..), breaks = breaks) +
  scale_x_continuous(limits = c(0,1), name = "Proportion of controls exposed") +
  scale_y_continuous(limits = c(1,6), name = "Minimum OR to detect")
print(direct.label(p, list("far.from.others.borders", "calc.boxes",
  "enlarge.box", hjust = 1, vjust = 1, box.color = NA,
  fill = "transparent", "draw.rects")))

## End(Not run)

## EXAMPLE 6:
## From page 1164 of Dupont (1988). A matched case control study is to be
## carried out to quantify the association between exposure A and an outcome B.
## Assume the prevalence of exposure in controls is 0.60 and the
## correlation between case and control exposures for matched pairs (rho) is
## 0.20 (moderate). Assuming an equal number of cases and controls, how many
## subjects need to be enrolled into the study to detect an odds ratio of 3.0
## with 0.80 power using a two-sided 0.05 test?
epi.sscc(OR = 3.0, p0 = 0.60, n = NA, power = 0.80, r = 1, rho.cc = 0.2,
  design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95,
  method = "matched", fleiss = FALSE)

## A total of 162 subjects need to be enrolled in the study: 81 cases and
## 81 controls.

## How many cases and controls are required if we select three
## controls per case?
epi.sscc(OR = 3.0, p0 = 0.60, n = NA, power = 0.80, r = 3, rho.cc = 0.2,
  design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95,
  method = "matched", fleiss = FALSE)

## A total of 204 subjects need to be enrolled in the study: 51 cases and
## 153 controls.
Description

Sample size to estimate a binary outcome using one-stage cluster sampling.

Usage

```r
epi.ssclus1estb(b, Py, epsilon, error = "relative",
    rho, nfractional = FALSE, conf.level = 0.95)
```

Arguments

- **b**: scalar integer or vector of length two, the number of individual listing units in each cluster to be sampled. See details, below.
- **Py**: scalar number, an estimate of the unknown population proportion.
- **epsilon**: scalar number, the maximum difference between the estimate and the unknown population value expressed in absolute or relative terms.
- **error**: character string. Options are `absolute` for absolute error and `relative` for relative error.
- **rho**: scalar number, the intracluster correlation.
- **nfractional**: logical, return fractional sample size.
- **conf.level**: scalar, defining the level of confidence in the computed result.

Details

- **b** as a scalar integer represents the total number of individual listing units from each cluster to be sampled. If **b** is a vector of length two the first element represents the mean number of individual listing units to be sampled from each cluster and the second element represents the standard deviation of the number of individual listing units to be sampled from each cluster.
- At least 25 primary sampling units are recommended for one-stage cluster sampling designs. If less than 25 clusters are returned by the function a warning is issued.

Value

A list containing the following:

- **n.psu**: the total number of primary sampling units (clusters) to be sampled for the specified level of confidence and relative error.
- **n.ssu**: the total number of secondary sampling units to be sampled for the specified level of confidence and relative error.
- **DEF**: the design effect.
- **rho**: the intracluster correlation, as entered by the user.

References

Examples

## EXAMPLE 1:

An aid project has distributed cook stoves in a single province in a
resource-poor country. At the end of three years, the donors would like
to know what proportion of households are still using their donated
stove. A cross-sectional study is planned where villages in the province
will be sampled and all households (approximately 75 per village) will be
visited to determine whether or not the donated stove is still in use.
A pilot study of the prevalence of stove usage in five villages
showed that 0.46 of householders were still using their stove. The
intracluster correlation for a study of this type is unknown, but thought
to be relatively high.

If the donor wanted to be 90% confident that the survey estimate of stove
usage was within 10% of the true population value, how many villages
(i.e. clusters) would need to be sampled?

epi.sscluslестb(b = 75, Py = 0.46, epsilon = 0.10, error = "relative",
rho = 0.20, nfractional = FALSE, conf.level = 0.90)

A total of 67 villages need to be sampled to meet the specifications
of this study.

Now imagine the situation where the number of households per village
varies. We are told that the average number of households per village is
75 with the 0.025 quartile 40 households and the 0.975 quartile 180
households. The expected standard deviation of the number of households
per village is (180 - 40) / 4 = 35. How many villages need to be sampled?

epi.sscluslестb(b = c(75,35), Py = 0.46, epsilon = 0.10, error = "relative",
rho = 0.20, nfractional = FALSE, conf.level = 0.90)

A total of 81 villages need to be sampled to meet the specifications
of this study.

epi.sscluslестc

Sample size to estimate a continuous outcome using one-stage cluster
sampling

Description

Sample size to estimate a continuous outcome using one-stage cluster sampling.

Usage

epi.sscluslестc(b, N, xbar, xsigma, epsilon, error = "relative",
rho, nfractional = FALSE, conf.level = 0.95)
Arguments

- **b**
  - scalar integer or vector of length two, the number of individual listing units in each cluster to be sampled. See details, below.

- **N**
  - scalar integer, representing the total number of individual listing units in the population.

- **xbar**
  - scalar number, the expected mean of the continuous variable to be estimated.

- **xsigma**
  - scalar number, the expected standard deviation of the continuous variable to be estimated.

- **epsilon**
  - scalar number, the maximum difference between the estimate and the unknown population value expressed in absolute or relative terms.

- **error**
  - character string. Options are **absolute** for absolute error and **relative** for relative error.

- **rho**
  - scalar number, the intracluster correlation.

- **nfractional**
  - logical, return fractional sample size.

- **conf.level**
  - scalar number, the level of confidence in the computed result.

Details

- **b** as a scalar integer represents the total number of individual listing units from each cluster to be sampled. If **b** is a vector of length two the first element represents the mean number of individual listing units to be sampled from each cluster and the second element represents the standard deviation of the number of individual listing units to be sampled from each cluster.

Value

A list containing the following:

- **n.psu**
  - the total number of primary sampling units (clusters) to be sampled for the specified level of confidence and relative error.

- **n.ssu**
  - the total number of secondary sampling units to be sampled for the specified level of confidence and relative error.

- **DEF**
  - the design effect.

- **rho**
  - the intracluster correlation, as entered by the user.

References


Examples

```r
## EXAMPLE 1:
## A survey to estimate the average number of residents over 75 years of
## age that require the services of a nurse in a given retirement village is
## to be carried out using a one-stage cluster sampling strategy.
## There are five housing complexes in the village with 25 residents in each.
## We expect that there might be an average of 34 residents meeting this
## criteria (SD 5.5). We would like the estimated sample size to provide us
## with an estimate that is within 10% of the true value. Previous studies
## report an intracluster correlation for the number of residents requiring the
## services of a nurse in this retirement village housing complexes to
## be 0.10. How many housing complexes (clusters) should be sampled?

epi.ssclus1estc(b = 25, N = 5 * 25, xbar = 34, xsigma = 5.5,
          epsilon = 0.10, error = "relative", rho = 0.10, nfractional = FALSE,
          conf.level = 0.95)
```

>> A total of 2 housing complexes need to be sampled to meet the specifications
>> of this study.

### epi.ssclus2estb

**Number of clusters to be sampled to estimate a binary outcome using**
**two-stage cluster sampling**

**Description**

Number of clusters to be sampled to estimate a binary outcome using two-stage cluster sampling.

**Usage**

```r
epi.ssclus2estb(b, Py, epsilon, error = "relative",
          rho, nfractional = FALSE, conf.level = 0.95)
```

**Arguments**

- `b` scalar integer or vector of length two, the number of individual listing units in each cluster to be sampled. See details, below.
- `Py` scalar number, an estimate of the unknown population proportion.
- `epsilon` scalar number, the maximum difference between the estimate and the unknown population value expressed in absolute or relative terms.
- `error` character string. Options are absolute for absolute error and relative for relative error.
- `rho` scalar number, the intracluster correlation.
- `nfractional` logical, return fractional sample size.
- `conf.level` scalar, defining the level of confidence in the computed result.
Details

b as a scalar integer represents the total number of individual listing units from each cluster to be sampled. If b is a vector of length two the first element represents the mean number of individual listing units to be sampled from each cluster and the second element represents the standard deviation of the number of individual listing units to be sampled from each cluster.

The methodology used in this function follows closely the approach described by Bennett et al. (1991). At least 25 primary sampling units are recommended for two-stage cluster sampling designs. If less than 25 clusters are returned by the function a warning is issued.

As a rule of thumb, around 30 clusters will provide good estimates of the true population value with an acceptable level of precision (Binkin et al. 1992) when: (1) the true population value is between 10% and 90%; and (2) the desired absolute error is around 5%. For a fixed number of individuals selected per cluster (e.g. 10 individuals per cluster or 30 individuals per cluster), collecting information on more than 30 clusters can improve the precision of the final population estimate, however, beyond around 60 clusters the improvement in precision is minimal.

Value

A list containing the following:

- n.psu
- n.ssu
- DEF
- rho

n.psu the total number of primary sampling units (clusters) to be sampled for the specified level of confidence and relative error.

n.ssu the total number of secondary sampling units to be sampled for the specified level of confidence and relative error.

DEF the design effect.

rho the intracluster correlation, as entered by the user.

References


Examples

```
## EXAMPLE 1 (from Bennett et al. 1991 p 102):
## We intend to conduct a cross-sectional study to determine the prevalence
## of disease X in a given country. The expected prevalence of disease is
## thought to be around 20%. Previous studies report an intracluster
## correlation for this disease to be 0.02. Suppose that we want to be 95%
## certain that our estimate of the prevalence of disease is within 5% of
## the true population value and that we intend to sample 20 individuals per
## cluster. How many clusters should be sampled to meet the requirements of the
## study?

epi.ssclus2estb(b = 20, Py = 0.20, epsilon = 0.05, error = "absolute",
```
rho = 0.02, nfractional = FALSE, conf.level = 0.95)

## A total of 17 clusters need to be sampled to meet the specifications
## of this study. epi.ssclus2estb returns a warning message that the number of
## clusters is less than 25.

epi.ssclus2estc

Number of clusters to be sampled to estimate a continuous outcome
using two-stage cluster sampling

Description

Number of clusters to be sampled to estimate a continuous outcome using two-stage cluster sam-
pling.

Usage

epi.ssclus2estc(b, N, xbar, xsigma, epsilon, error = "relative",
               rho, nfractional = FALSE, conf.level = 0.95)

Arguments

b         scalar integer or vector of length two, the number of individual listing units in
each cluster to be sampled. See details, below.
N         scalar integer, representing the total number of individual listing units in the
population.
xbar      scalar number, the expected mean of the continuous variable to be estimated.
xsigma   scalar number, the expected standard deviation of the continuous variable to be
estimated.
epsilon   scalar number, the maximum difference between the estimate and the unknown
population value expressed in absolute or relative terms.
error     character string. Options are absolute for absolute error and relative for
relative error.
rho       scalar number, the intracluster correlation.
nfractional logical, return fractional sample size.
conf.level scalar number, the level of confidence in the computed result.

Details

b as a scalar integer represents the total number of individual listing units from each cluster to be
sampled. If b is a vector of length two the first element represents the mean number of individual
listing units to be sampled from each cluster and the second element represents the standard
deviation of the number of individual listing units to be sampled from each cluster.
Value

A list containing the following:

- **n.psu**: the total number of primary sampling units (clusters) to be sampled for the specified level of confidence and relative error.
- **n.ssu**: the total number of secondary sampling units to be sampled for the specified level of confidence and relative error.
- **DEF**: the design effect.
- **rho**: the intracluster correlation, as entered by the user.

References


Examples

### EXAMPLE 1 (from Levy and Lemeshow p 292):

We intend to conduct a survey of nurse practitioners to estimate the average number of patients seen by each nurse. There are five health centres in the study area, each with three nurses. We intend to sample two nurses from each health centre. We would like to be 95% confident that our estimate is within 30% of the true population value. We expect that the mean number of patients seen at the health centre level is 84 (var 567) and the mean number of patients seen at the nurse level is 28 (var 160). Previous studies report an intracluster correlation for the number of patients seen per nurse to be 0.02. How many health centres should be sampled?

```r
epi.ssclus2estc(b = 2, N = 15, xbar = 28, xsigma = sqrt(160),
epsilon = 0.30, error = "relative", rho = 0.02, nfractional = FALSE,
conf.level = 0.95)
```

## A total of 3 health centres need to be sampled to meet the specifications of this study.

---

epi.sscohortc

Sample size, power or minimum detectable incidence risk ratio for a cohort study using individual count data

Description

Sample size, power or minimum detectable incidence risk ratio for a cohort study using individual count data.
### Usage

```r
epi.sscohortc(irexp1, irexp0, pexp = NA, n = NA, power = 0.80, r = 1,
N, design = 1, sided.test = 2, finite.correction = FALSE,
nfractional = FALSE, conf.level = 0.95)
```

### Arguments

- **irexp1**: the expected incidence risk of the outcome in the exposed group (0 to 1).
- **irexp0**: the expected incidence risk of the outcome in the non-exposed group (0 to 1).
- **pexp**: the expected prevalence of exposure to the hypothesised risk factor in the population (0 to 1).
- **n**: scalar, defining the total number of subjects in the study (i.e. the number in both the exposed and unexposed groups).
- **power**: scalar, the required study power.
- **r**: scalar, the number in the exposed group divided by the number in the unexposed group.
- **N**: scalar, the estimated number of individuals in the population.
- **design**: scalar, the estimated design effect.
- **sided.test**: use a one- or two-sided test? Use a two-sided test if you wish to evaluate whether or not the outcome incidence risk in the exposed group is greater than or less than the outcome incidence risk in the unexposed group. Use a one-sided test to evaluate whether or not the outcome incidence risk in the exposed group is greater than the outcome incidence risk in the unexposed group.
- **finite.correction**: logical, apply a finite correction factor?
- **nfractional**: logical, return fractional sample size.
- **conf.level**: scalar, defining the level of confidence in the computed result.

### Details

The methodology in this function follows the approach described in Chapter 8 of Woodward (2005), pp. 381 - 426.

### Value

A list containing the following:

- **n.total**: the total number of subjects required for the specified level of confidence and power, respecting the requirement for \( r \) times as many individuals in the exposed (treatment) group compared with the non-exposed (control) group.
- **n.exp1**: the total number of subjects in the exposed (treatment) group for the specified level of confidence and power, respecting the requirement for \( r \) times as many individuals in the exposed (treatment) group compared with the non-exposed (control) group.
n.exp0  the total number of subjects in the non-exposed (control) group for the specified level of confidence and power, respecting the requirement for \( r \) times as many individuals in the exposed (treatment) group compared with the non-exposed (control) group.

power  the power of the study given the number of study subjects, the expected effect size and level of confidence.

irr  the incidence risk of the outcome in the exposed group divided by the incidence risk of the outcome in the unexposed group (the incidence risk ratio).

or  the odds of the outcome in the exposed group divided by the odds of the outcome in the unexposed group (the odds ratio).

Note  The power of a study is its ability to demonstrate the presence of an association, given that an association actually exists.

Values need to be entered for \( i_{\text{exp}0} \), \( p_{\text{exp}} \), \( n \), and power to return a value for \( \text{irr} \). In this situation, the lower value of \( \text{irr} \) represents the maximum detectable incidence risk ratio that is less than 1; the upper value of \( \text{irr} \) represents the minimum detectable incidence risk ratio greater than 1. A value for \( p_{\text{exp}} \) doesn’t need to be entered if you want to calculate sample size or study power.

When calculating study power or minimum detectable incidence risk ratio when \texttt{finite.correction} = TRUE the function takes the values of \( n \) and \( N \) entered by the user and back-calculates a value of \( n \) assuming an infinite population. Values for power, \( \text{irr} \) and \( \text{or} \) are then returned, assuming the back-calculated value of \( n \) is equivalent to the value of \( n \) entered by the user.

References


Examples

```r
## EXAMPLE 1 (from Woodward 2005 p. 406):
## A cohort study of smoking and coronary heart disease (CHD) in middle aged men
## is planned. A sample of men will be selected at random from the population
## and those that agree to participate will be asked to complete a
## questionnaire. The follow-up period will be 5 years. The investigators would
## like to be 0.90 sure of being able to detect when the risk ratio of CHD
## is 1.4 for smokers, using a 0.05 significance test. Previous evidence
## suggests that the incidence risk of death in non-smokers is 413 per
## 100,000 per year. Assuming equal numbers of smokers and non-smokers are
## sampled, how many men should be sampled overall?

irexp1 = 1.4 * (5 * 413) / 100000; irexp0 = (5 * 413) / 100000
epi.sscohortc(irexp1 = irexp1, irexp0 = irexp0, pexp = NA, n = NA,
  power = 0.90, r = 1, N = NA, design = 1, sided.test = 1,
  finite.correction = FALSE, nfractional = FALSE, conf.level = 0.95)
```
## A total of 12,130 men need to be sampled (6065 smokers and 6065 non-smokers).

## EXAMPLE 2 (from Woodward 2005 p. 406):
Say, for example, we are only able to enrol 5000 subjects into the study described above. What is the minimum and maximum detectable risk ratio?

\[
\text{irexp0} = \frac{5 \times 413}{100000}
\]

epi.sscohortc(irexp1 = NA, irexp0 = irexp0, pexp = NA, n = 5000,
    power = 0.90, r = 1, N = NA, design = 1, sided.test = 1,
    finite.correction = FALSE, nfractional = FALSE, conf.level = 0.95)

## The minimum detectable risk ratio >1 is 1.65. The maximum detectable risk ratio <1 is 0.50.

## EXAMPLE 3:
A study is to be carried out to assess the effect of a new treatment for anoestrus in dairy cattle. What is the required sample size if we expect the proportion of cows responding in the treatment (exposed) group to be 0.30 and the proportion of cows responding in the control (unexposed) group to be 0.15? The required power for this study is 0.80 using a two-sided 0.05 test.

epi.sscohortc(irexp1 = 0.30, irexp0 = 0.15, pexp = NA, n = NA,
    power = 0.80, r = 1, N = NA, design = 1, sided.test = 2,
    finite.correction = FALSE, nfractional = FALSE, conf.level = 0.95)

## A total of 242 cows are required: 121 in the treatment (exposed) group and 121 in the control (unexposed) group.

Assume now that this study is going to be carried out using animals from a number of herds. What is the required sample size when you account for the observation that response to treatment is likely to cluster within herds?

For the exercise, assume that the intra-cluster correlation coefficient (the rate of homogeneity, \(\rho\)) for this treatment is 0.05 and the average number of cows sampled per herd will be 30.

Calculate the design effect, given \(\rho = (\text{design} - 1) / (\text{nbar} - 1)\), where \(\text{nbar}\) equals the average number of individuals per cluster:

\[
\text{design} \leftarrow 0.05 \times (30 - 1) + 1; \text{design}
\]

The design effect is 2.45.

epi.sscohortc(irexp1 = 0.30, irexp0 = 0.15, pexp = NA, n = NA,
    power = 0.80, r = 1, N = NA, design = design, sided.test = 2,
    finite.correction = FALSE, nfractional = FALSE, conf.level = 0.95)

## A total of 592 cows are required for this study: 296 in the treatment group and 296 in the control group.
Sample size, power or minimum detectable incidence rate ratio for a cohort study using person or animal time data

Usage

epi.sscohortt(irexp1, irexp0, FT = NA, n, power, r = 1, design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95)

Arguments

irexp1 the expected incidence rate of the outcome in the exposed group (0 to 1).
irexp0 the expected incidence rate of the outcome in the non-exposed group (0 to 1).
FT the follow-up period (in years) for the study.
n scalar, defining the total number of subjects in the study (i.e. the number in both the exposed and unexposed groups).
power scalar, the required study power.
r scalar, the number in the exposed group divided by the number in the unexposed group.
design scalar, the estimated design effect.
sided.test use a one- or two-sided test? Use a two-sided test if you wish to evaluate whether or not the outcome incidence rate in the exposed group is greater than or less than the outcome incidence rate in the unexposed group. Use a one-sided test to evaluate whether or not the outcome incidence rate in the exposed group is greater than the outcome incidence rate in the unexposed group.
nfractional logical, return fractional sample size.
conf.level scalar, defining the level of confidence in the computed result.

Details

The methodology in this function follows the approach described in Lwanga and Lemeshow (1991).

Value

A list containing the following:

n.total the total number of subjects required for the specified level of confidence and power, respecting the requirement for \( r \) times as many individuals in the treatment group compared with the control group.
the total number of subjects in the treatment group for the specified level of confidence and power, respecting the requirement for $r$ times as many individuals in the treatment group compared with the control group.

$n.\exp0$ the total number of subjects in the control group for the specified level of confidence and power, respecting the requirement for $r$ times as many individuals in the treatment group compared with the control group.

$power$ the power of the study given the number of study subjects, the expected effect size and level of confidence.

$irr$ the incidence rate of the outcome in the exposed group divided by the incidence rate in the unexposed group (the incidence rate ratio).

Note

The power of a study is its ability to demonstrate the presence of an association, given that an association actually exists.

Values need to be entered for $irexp0$, $n$, and $power$ to return a value for $irr$. In this situation, the lower value of $irr$ represents the maximum detectable incidence rate ratio that is less than 1; the upper value of $irr$ represents the minimum detectable incidence rate ratio greater than 1.

See the documentation for epi.sscohortc for an example using the design facility implemented in this function.

References


Examples

```r
## EXAMPLE 1 (from Lwanga and Lemeshow 1991 p. 19):
## As part of a study of the long-term effect of noise on workers in a
## particularly noisy industry, it is planned to follow up a cohort of people
## who were recruited into the industry during a given period of time and to
## compare them with a similar cohort of individuals working in a much
## quieter industry. Subjects will be followed up for the rest of their lives or
## until their hearing is impaired. The results of a previous small-scale survey
## suggest that the annual incidence rate of hearing impairment in the noisy
## industry may be as high as 25%. How many people should be followed up
## in each of the groups (which are to be of equal size) to test the hypothesis
## that the incidence rates for hearing impairment in the two groups are the
## same, at the 5% level of significance and with a power of 80%? The
## alternative hypothesis is that the annual incidence rate for hearing
## impairment in the quieter industry is not more than the national average of
## about 10% (for people in the same age range), whereas in the noisy
## industry it differs from this.

## An annual incidence rate of 25% is equivalent to 25 cases of hearing
## impairment per 100 individuals per year.
```
Sample size, power and minimum detectable risk ratio when comparing binary outcomes.

**Usage**

```r
epi.sscompb(treat, control, n, power, r = 1, design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95)
```

**Arguments**

- **treat**: the expected proportion for the treatment group (see below).
- **control**: the expected proportion for the control group (see below).
- **n**: scalar, defining the total number of subjects in the study (i.e. the number in the treatment plus the number in the control group).
- **power**: scalar, the required study power.
- **r**: scalar, the number in the treatment group divided by the number in the control group.
- **design**: scalar, the estimated design effect.
- **sided.test**: use a one- or two-sided test? Use a two-sided test if you wish to evaluate whether or not the outcome proportion in the exposed (treatment) group is greater than or less than the outcome proportion in the unexposed (control) group. Use a one-sided test to evaluate whether or not the outcome proportion in the exposed (treatment) group is greater than the outcome proportion in the unexposed (control) group.
epi.sscompb

nfractional logical, return fractional sample size.
conf.level scalar, defining the level of confidence in the computed result.

details

The methodology in this function follows the approach described in Chapter 8 of Woodward (2005).

With this function it is assumed that one of the two proportions is known and we want to test the null hypothesis that the second proportion is equal to the first. Users are referred to the epi.sscohortc function which relates to the two-sample problem where neither proportion is known (or assumed, at least).

Because there is much more uncertainty in the two sample problem where neither proportion is known, epi.sscohortc returns much larger sample size estimates. This function (epi.sscompb) should be used in particular situations such as when a politician claims that at least 90% of the population use seatbelts and we want to see if the data supports this claim.

value

A list containing the following:

n.total the total number of subjects required for the specified level of confidence and power, respecting the requirement for \( r \) times as many individuals in the treatment group compared with the control group.
n.treat the total number of subjects in the treatment group for the specified level of confidence and power, respecting the requirement for \( r \) times as many individuals in the treatment group compared with the control group.
n.control the total number of subjects in the control group for the specified level of confidence and power, respecting the requirement for \( r \) times as many individuals in the treatment group compared with the control group.
power the power of the study given the number of study subjects, the expected effect size and level of confidence.
lambda the proportion in the treatment group divided by the proportion in the control group (a risk ratio).

note

The power of a study is its ability to demonstrate the presence of an association, given that an association actually exists.

Values need to be entered for control, n, and power to return a value for lambda. In this situation, the lower value of lambda represents the maximum detectable risk ratio that is less than 1; the upper value of lambda represents the minimum detectable risk ratio greater than 1.

See the documentation for epi.sscohortc for an example using the design facility implemented in this function.
References


Examples

## EXAMPLE 1 (from Woodward 2005 pp. 403 - 404):
## A government initiative has decided to reduce the prevalence of male smoking to, at most, 30%. A sample survey is planned to test, at the 0.05 level, the hypothesis that the percentage of smokers in the male population is 30% against the one-sided alternative that it is greater. The survey should be able to find a prevalence of 32%, when it is true, with 0.90 power. How many men need to be sampled?

epi.sscompb(treat = 0.30, control = 0.32, n = NA, power = 0.90, r = 1, design = 1, sided.test = 1, nfractional = FALSE, conf.level = 0.95)

## A total of 18,316 men should be sampled: 9158 in the treatment group and 9158 in the control group. The risk ratio (that is, the prevalence of smoking in males post government initiative divided by the prevalence of smoking in males pre inititative is 0.94.

## EXAMPLE 2:
## If we sample only 10,000 men (5000 in the treatment group and 5000 in the control group) what is the maximum detectable risk ratio that is less than 1?

epi.sscompb(treat = NA, control = 0.32, n = 10000, power = 0.90, r = 1, design = 1, sided.test = 1, nfractional = FALSE, conf.level = 0.95)

## If we sample only 10,000 men the maximum detectable risk ratio will be 0.91.

epi.sscompc Sample size, power and minimum detectable difference when comparing continuous outcomes

Description

Sample size, power and minimum detectable difference when comparing continuous outcomes.

Usage

epi.sscompc(treat, control, n, sigma, power, r = 1, design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95)
Arguments

treat the expected value for the treatment group (see below).
control the expected value for the control group (see below).
n scalar, defining the total number of subjects in the study (i.e. the number in the treatment and control group).
sigma the expected standard deviation of the variable of interest for both treatment and control groups.
power scalar, the required study power.
r scalar, the number in the treatment group divided by the number in the control group.
design scalar, the estimated design effect.
sided.test use a one- or two-sided test? Use a two-sided test if you wish to evaluate whether or not the outcome in the exposed (treatment) group is greater than or less than the outcome in the unexposed (control) group. Use a one-sided test to evaluate whether or not the outcome in the exposed (treatment) group is greater than the outcome in the unexposed (control) group.
nfractional logical, return fractional sample size.
conf.level scalar, defining the level of confidence in the computed result.

Details

The methodology in this function follows the approach described in Chapter 8 of Woodward (2005), pp. 381 - 426.

Value

A list containing the following:

n.total the total number of subjects required for the specified level of confidence and power, respecting the requirement for r times as many individuals in the treatment group compared with the control group.
n.treat the total number of subjects in the treatment group for the specified level of confidence and power, respecting the requirement for r times as many individuals in the treatment group compared with the control group.
n.control the total number of subjects in the control group for the specified level of confidence and power, respecting the requirement for r times as many individuals in the treatment group compared with the control group.
power the power of the study given the number of study subjects, the expected effect size and level of confidence.
delta the minimum detectable difference given the specified level of confidence and power.
Note

The power of a study is its ability to demonstrate the presence of an association, given that an association actually exists.

A detailed description of sample size calculations for case-control studies (with numerous worked examples, many of them reproduced below) is provided by Woodward (2005), pages 381 to 426.

See the documentation for `epi.sscohortc` for an example using the design facility implemented in this function.

References


Examples

```r
## EXAMPLE 1 (from Woodward 2005 p. 399):
## Supposed we wish to test, at the 5% level of significance, the hypothesis
## that cholesterol means in a population are equal in two study years against
## the one-sided alternative that the mean is higher in the second of the
## two years. Suppose that equal sized samples will be taken in each year,
## but that these will not necessarily be from the same individuals (i.e. the
## two samples are drawn independently). Our test is to have a power of 0.95
## at detecting a difference of 0.5 mmol/L. The standard deviation of serum
## cholesterol in humans is assumed to be 1.4 mmol/L.

epi.sscompc(treat = 5, control = 4.5, n = NA, sigma = 1.4, power = 0.95,
            r = 1, design = 1, sided.test = 1, nfractional = FALSE, conf.level = 0.95)
## To satisfy the study requirements 340 individuals need to be tested: 170 in
## the first year and 170 in the second year.

## EXAMPLE 2 (from Woodward 2005 pp. 399 - 400):
## Women taking oral contraceptives sometimes experience anaemia due to
## impaired iron absorption. A study is planned to compare the use of iron
## tablets against a course of placebos. Oral contraceptive users are
## randomly allocated to one of the two treatment groups and mean serum
## iron concentration compared after 6 months. Data from previous studies
## indicates that the standard deviation of the increase in iron
## concentration will be around 4 micrograms% over a 6-month period.
## The average increase in serum iron concentration without supplements is
## also thought to be 4 micrograms%. The investigators want to be 90% sure
## of detecting when the supplement doubles the serum iron concentration using
## a two-sided 5% significance test. It is decided to allocate 4 times as many
## women to the treatment group so as to obtain a better estimate of its effect.
## How many women should be enrolled in this study?

epi.sscompc(treat = 8, control = 4, n = NA, sigma = 4, power = 0.90,
            r = 1, design = 1, sided.test = 1, nfractional = FALSE, conf.level = 0.95)
```
r = 4, design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95)

## The estimated sample size is 70. We allocate 70/5 = 14 women to the
## placebo group and four times as many (56) to the iron treatment group.

desc
Sample size, power and minimum detectable hazard when comparing
time to event

Description
Sample size, power and minimum detectable hazard when comparing time to event.

Usage
epi.sscomps(treat, control, n, power, r = 1, design = 1,
sided.test = 2, nfractional = FALSE, conf.level = 0.95)

Arguments
treat the expected value for the treatment group (see below).
control the expected value for the control group (see below).
n scalar, defining the total number of subjects in the study (i.e. the number in the
treatment and control group).
power scalar, the required study power.
r scalar, the number in the treatment group divided by the number in the control
group. This argument is ignored when method = "proportions".
design scalar, the estimated design effect.
sided.test use a one- or two-sided test? Use a two-sided test if you wish to evaluate whether
or not the outcome hazard in the exposed (treatment) group is greater than or less
than the outcome hazard in the unexposed (control) group. Use a one-sided test
to evaluate whether or not the outcome hazard in the exposed (treatment) group
is greater than the outcome hazard in the unexposed (control) group.
nfractional logical, return fractional sample size.
conf.level scalar, defining the level of confidence in the computed result.

Details
The argument treat is the proportion of treated subjects that will have not experienced the event
of interest at the end of the study period and control is the proportion of control subjects that
will have not experienced the event of interest at the end of the study period. See Therneau and
Grambsch pp 61 - 65.
Value

A list containing one or more of the following:

- **n.crude**: the crude estimated total number of events required for the specified level of confidence and power.
- **n.total**: the total estimated number of events required for the specified level of confidence and power, respecting the requirement for \( r \) times as many events in the treatment group compared with the control group.
- **hazard**: the minimum detectable hazard ratio >1 and the maximum detectable hazard ratio <1.
- **power**: the power of the study given the number of events, the expected hazard ratio and level of confidence.

Note

The power of a study is its ability to demonstrate the presence of an association, given that an association actually exists.

See the documentation for `epi.sscohortc` for an example using the design facility implemented in this function.

References


Examples

```r
## EXAMPLE 1 (from Therneau and Grambsch 2000 p. 63):
## The 5-year survival probability of patients receiving a standard treatment
## is 0.30 and we anticipate that a new treatment will increase it to 0.45.
## Assume that a study will use a two-sided test at the 0.05 level with 0.90
## power to detect this difference. How many events are required?
epi.sscomps(treat = 0.45, control = 0.30, n = NA, power = 0.90,
           r = 1, design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95)
## A total of 250 events are required. Assuming one event per individual,
## assign 125 individuals to the treatment group and 125 to the control group.

## EXAMPLE 2 (from Therneau and Grambsch 2000 p. 63):
## What is the minimum detectable hazard in a study involving 500 subjects where
## the treatment to control ratio is 1:1, assuming a power of 0.90 and a
## 2-sided test at the 0.05 level?
epi.sscomps(treat = NA, control = NA, n = 500, power = 0.90,
           r = 1, design = 1, sided.test = 2, nfractional = FALSE, conf.level = 0.95)
```
## Assuming treatment increases time to event (compared with controls), the
## minimum detectable hazard of a study involving 500 subjects (250 in the
treatment group and 250 in the controls) is 1.33.

### epi.ssdetect

**Sample size to detect an event**

**Description**

Sample size to detect at least one event (e.g. a disease-positive individual) in a population. The method adjusts sample size estimates on the basis of test sensitivity and can account for series and parallel test interpretation.

**Usage**

```r
epi.ssdetect(N, prev, se, sp, interpretation = "series", covar = c(0,0),
finite.correction = TRUE, nfractional = FALSE, conf.level = 0.95)
```

**Arguments**

- `N`: a vector of length one or two defining the size of the population. The first element of the vector defines the number of clusters, the second element defining the mean number of sampling units per cluster.
- `prev`: a vector of length one or two defining the prevalence of disease in the population. The first element of the vector defines the between-cluster prevalence, the second element defines the within-cluster prevalence.
- `se`: a vector of length one or two defining the sensitivity of the test(s) used.
- `sp`: a vector of length one or two defining the specificity of the test(s) used.
- `interpretation`: a character string indicating how test results should be interpreted. Options are `series` or `parallel`.
- `covar`: a vector of length two defining the covariance between test results for disease positive and disease negative groups. The first element of the vector is the covariance between test results for disease positive subjects. The second element of the vector is the covariance between test results for disease negative subjects. Use `covar = c(0,0)` (the default) if these values are not known.
- `finite.correction`: logical, apply finite correction? See details, below.
- `nfractional`: logical, return fractional sample size.
- `conf.level`: scalar, defining the level of confidence in the computed result.

**Value**

A list containing the following:

- `performance`: The sensitivity and specificity of the testing strategy.
- `sample.size`: The number of clusters, units, and total number of units to be sampled.
Note

Sample size calculations are carried out using the binomial distribution and an approximation of the hypergeometric distribution (MacDiarmid 1988). Because the hypergeometric distribution takes into account the size of the population being sampled, `finite.correction = TRUE` is only applied to the binomial sample size estimates.

Define `se1` and `se2` as the sensitivity for the first and second test, `sp1` and `sp2` as the specificity for the first and second test, `p111` as the proportion of disease-positive subjects with a positive test result to both tests and `p000` as the proportion of disease-negative subjects with a negative test result to both tests. The covariance between test results for the disease-positive group is $p111 - se1 * se2$. The covariance between test results for the disease-negative group is $p000 - sp1 * sp2$.

References


Examples

```r
## EXAMPLE 1:
## We would like to confirm the absence of disease in a single 1000-cow
dairy herd. We expect the prevalence of disease in the herd to be 5%. We intend to use a single test with a sensitivity of 0.90 and a
## specificity of 1.00. How many samples should we take to be 95% certain that, if all tests are negative, the disease is not present?

epi.ssdetect(N = 1000, prev = 0.05, se = 0.90, sp = 1.00, interpretation = "series",
finite.correction = TRUE, nfractional = FALSE, conf.level = 0.95)

## Using the hypergeometric distribution, we need to sample 65 cows.

## EXAMPLE 2:
## We would like to confirm the absence of disease in a study area. If the
disease is present we expect the between-herd prevalence to be 8% and the
## within-herd prevalence to be 5%. We intend to use two tests: the first has
## a sensitivity and specificity of 0.90 and 0.80, respectively. The second
## has a sensitivity and specificity of 0.95 and 0.85, respectively. The two
## tests will be interpreted in parallel. How many herds and cows within herds
## should we sample to be 95% certain that the disease is not present in the
## study area if all tests are negative? There area is comprised of
## approximately 5000 herds and the average number of cows per herd is 100.

epi.ssdetect(N = c(5000, 100), prev = c(0.08, 0.05), se = c(0.90, 0.95),
sp = c(0.80, 0.85), interpretation = "parallel", covar = c(0,0),
finite.correction = TRUE, nfractional = FALSE, conf.level = 0.95)
```
## We need to sample 46 cows from 40 herds (a total of 1840 samples).
## The sensitivity of this testing regime is 99%. The specificity of this
## testing regime is 68%.

## EXAMPLE 3:
## You want to document the absence of Mycoplasma from a 200-sow pig herd.
## Based on your experience and the literature, a minimum of 20% of sows
## would have seroconverted if Mycoplasma were present in the herd. How many
## sows do you need to sample?

```r
epi.ssdetect(N = 200, prev = 0.20, se = 1.00, sp = 1.00, interpretation =
"series", covar = c(0,0), finite.correction = TRUE, nfractional = FALSE,
conf.level = 0.95)
```

## If you test 15 sows and all test negative you can state that you are 95%
## confident that the prevalence rate of Mycoplasma in the herd is less than
## 20%.

---

### epi.ssdxtest

**Sample size to validate a diagnostic test in the absence of a gold standard**

#### Description

Sample size to validate a diagnostic test in the absence of a gold standard.

#### Usage

```r
epi.ssdxtest(pi, se, sp, epsilon.api, epsilon.ase, epsilon.asp, epsilon.asesp,
r = 1, nfractional = FALSE, verbose = FALSE, conf.level = 0.95)
```

#### Arguments

- **pi**: vector of length two, the expected prevalence of the outcome of interest in the two populations (0 to 1), respectively.
- **se**: vector of length two, the expected diagnostic sensitivity of the first and second test (0 to 1), respectively.
- **sp**: vector of length two, the expected diagnostic specificity of the first and second test (0 to 1), respectively.
- **epsilon.api**: vector of length two, the absolute error for the prevalence of the outcome of interest in the first and second study populations.
- **epsilon.ase**: vector of length two, the absolute error for the sensitivity estimate of the first and second test.
- **epsilon.asp**: vector of length two, the absolute error for the specificity estimate of the first and second test.
epsilon.asesp vector of length two, the absolute error for the difference in the two sensitivity and specificity estimates.

r scalar, the required number in the second population divided by the number in the first population.

nfractional logical, return fractional sample size.

verbose logical, return sample size estimates for se, sp, and pi?

conf.level scalar, defining the level of confidence in the computed result.

Details

Hui and Walter (1980) describe an approach for estimating the sensitivity and specificity of a diagnostic test in the absence of a gold standard. Their method involves testing individuals from two populations with two conditionally independent diagnostic tests (neither of which is a gold standard). With such data, all six parameters of interest (two sensitivities, two specificities and two prevalences) can be estimated since there are six degrees of freedom available. The methodology in this function follows the sample size calculation methods described by Georgiadis et al (2005).

In their paper Georgiadis et al (2005) parameterise the uncertainty in the prevalence, sensitivity and specificity estimates in terms of the width of the confidence interval. For consistency with the other sample size calculation functions in epiR the amount of uncertainty in the prevalence, sensitivity and specificity is parameterised in absolute terms. Using this approach, if we set $se = c(0.80, 0.90)$ and $epsilon.ase = c(0.05, 0.10)$ the number of subjects to return an estimate of $se1$ that is between 0.75 and 0.85 and $se2$ that is between 0.80 and 1.0 will be returned.

Value

When verbose = TRUE a data frame listing the required sample size to estimate:

- $p1$ the prevalence of the outcome of interest in population 1.
- $p2$ the prevalence of the outcome of interest in population 2.
- $se1$ the sensitivity of the first diagnostic test.
- $se2$ the sensitivity of the second diagnostic test.
- $sp1$ the specificity of the first diagnostic test.
- $sp2$ the specificity of the second diagnostic test.
- $se1.se2$ the difference in the sensitivities of the two diagnostic tests.
- $sp1.sp2$ the difference in the specificities of the two diagnostic tests.

When verbose = FALSE a data frame listing the maximum of the sample size estimates listed when verbose = TRUE.

References


Examples

## EXAMPLE 1 (from Georgiadis et al. 2005, pp. 5):
## Nielsen et al. (2002) present data from the evaluation of a milk antibody ELISA and faecal culture for the diagnosis of Mycobacterium avium subsp. paratuberculosis infection in cattle. Because the ELISA detects antibodies and culture is based on isolation of the bacterium in faeces we can reasonably assume that the two tests are conditionally independent.

## How many cattle need to be sampled if we wanted to be 95% certain that our estimate of diagnostic sensitivity and specificity of the two tests is within 0.05 of the true population value assuming the number sampled in the second population divided by the number sampled in the first population is 0.817? The prevalence of Mycobacterium avium subsp. paratuberculosis is thought to be 0.093 and 0.204, respectively. Assume the sensitivity of the the ELISA and faecal culture is 0.349 and 0.534, respectively. Assume the specificity of the ELISA and faecal culture is 0.995 and 0.894, respectively.

epi.ssdxtest(pi = c(0.093, 0.204), se = c(0.349, 0.534), sp = c(0.995, 0.894),
epsilon.api = c(0.05, 0.05), epsilon.ase = c(0.05, 0.05),
epsilon.asp = c(0.05, 0.05), epsilon.asesp = c(0.05, 0.05),
r = 0.817, nfractional = FALSE, verbose = FALSE, conf.level = 0.95)

## A total of 63,887 cattle need to be sampled (35,161 from population 1 and 28,726 from population 2) to meet the requirements of the study.

---

epi.ssequb  
**Sample size for a parallel equivalence trial, binary outcome**

### Description
Sample size for a parallel equivalence trial, binary outcome.

### Usage
epi.ssequb(treat, control, delta, n, r = 1, power, nfractional = FALSE, alpha)

### Arguments
- **treat**: the expected proportion of successes in the treatment group.
- **control**: the expected proportion of successes in the control group.
- **delta**: the equivalence limit, expressed as the change in the outcome of interest that represents a clinically meaningful difference.
n scalar, the total number of study subjects in the trial.

r scalar, the number in the treatment group divided by the number in the control group.

power scalar, the required study power.

n_fractional logical, return fractional sample size.

alpha scalar, defining the desired alpha level.

Value

A list containing the following:

- n.total the total number of study subjects required.
- n.treat the required number of study subject in the treatment group.
- n.control the required number of study subject in the control group.
- power the specified or calculated study power.

Note

Consider a clinical trial comparing two groups, a standard treatment (s) and a new treatment (n). In each group, a proportion of subjects respond to the treatment: $P_s$ and $P_n$.

With a superiority trial we specify the maximum acceptable difference between $P_s$ and $P_n$ as $\delta$. The null hypothesis is $H_0 : P_n - P_s \leq \delta$ and the alternative hypothesis is $H_1 : P_n - P_s > \delta$.

An equivalence trial is used if you want to prove that two treatments produce the same clinical outcomes. With an equivalence trial, we specify the maximum acceptable difference between $P_n$ and $P_s$ as $\delta$. The null hypothesis is $H_0 : |P_n - P_s| \geq \delta$ and the alternative hypothesis is $H_1 : |P_n - P_s| < \delta$. In bioequivalence trials, a 90% confidence interval is often used. The value of the maximum acceptable difference $\delta$ is chosen so that a patient will not detect any change in effect when replacing the standard treatment with the new treatment.

With a non-inferiority trial we specify the maximum acceptable difference between $P_n$ and $P_s$ as $\delta$. The null hypothesis is $H_0 : P_s - P_n \geq \delta$ and the alternative hypothesis is $H_1 : P_s - P_n < \delta$.

The aim of a non-inferiority trial is show that a new treatment is not (much) inferior to a standard treatment. Showing non-inferiority can be of interest because: (a) it is often not ethically possible to do a placebo-controlled trial, (b) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints, but is safer, (c) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints, but is cheaper to produce or easier to administer, (d) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints in clinical trial, but compliance will be better outside the clinical trial and hence efficacy better outside the trial.

To summarise (adapted from Machin et al. 2009, page 105):

<table>
<thead>
<tr>
<th>Test for</th>
<th>Null hypothesis</th>
<th>Alt hypothesis</th>
<th>Type I</th>
<th>Type II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superiority</td>
<td>$H_0 : P_n - P_s \leq \delta$</td>
<td>$H_1 : P_n - P_s &gt; \delta$</td>
<td>2 sided, 5.0%</td>
<td>1 sided, 10 or 20%</td>
</tr>
<tr>
<td>Equivalence</td>
<td>$H_0 :</td>
<td>P_n - P_s</td>
<td>\geq \delta$</td>
<td>$H_1 :</td>
</tr>
<tr>
<td>Non-inferiority</td>
<td>$H_0 : P_n - P_s \geq \delta$</td>
<td>$H_1 : P_n - P_s &lt; \delta$</td>
<td>1 sided, 2.5%</td>
<td>1 sided, 10 or 20%</td>
</tr>
</tbody>
</table>
Superiority trial: $H_1$ is that the new treatment is better than the standard treatment.
Equivalence trial: $H_1$ is that the new treatment is not too different from the standard treatment.
Non-inferiority trial: $H_1$ is that the new treatment is not much worse than the standard treatment.

When calculating the power of a study, note that the variable $n$ refers to the total study size (that is, the number of subjects in the treatment group plus the number in the control group).

References


Examples

```r
## EXAMPLE 1 (from Machin, Campbell, Tan and Tan 2009 p. 113):
## Bennett, Dismukes, Duma et al. (1979) designed a clinical trial to test
## whether combination chemotherapy for a shorter period would be at least
## as good as conventional therapy for patients with cryptococcal meningitis.
## They recruited 39 patients to each treatment arm and wished to conclude
## that a difference of less than 20% in response rate between the treatments
## would indicate equivalence. Assuming a one-sided test size of 10%, a
## power of 80% and an overall response rate of 50%, what would be a
## realistic sample size if the trial were to be repeated?

epi.ssequb(treat = 0.50, control = 0.50, delta = 0.20, n = NA, r = 1,
           power = 0.80, nfractional = FALSE, alpha = 0.10)
```

## A total of 166 subjects need to be enrolled in the trial, 83 in the
## treatment group and 83 in the control group.

---

**Description**

Sample size for a parallel equivalence trial, continuous outcome.

**Usage**

```r
epi.ssequc(treat, control, sd, delta, n, r = 1, power, nfractional = FALSE,
           alpha)
```
Arguments

treat
the expected mean of the outcome of interest in the treatment group.

treat
the expected mean of the outcome of interest in the control group.

sd
the expected population standard deviation of the outcome of interest.

delta
the equivalence limit, expressed as the change in the outcome of interest that represents a clinically meaningful difference.

n
scalar, the total number of study subjects in the trial.

r
scalar, the number in the treatment group divided by the number in the control group.

power
scalar, the required study power.

n_fractional
logical, return fractional sample size.

alpha
scalar, defining the desired alpha level.

Value
A list containing the following:

  n.total
  the total number of study subjects required.

  n.treat
  the required number of study subject in the treatment group.

  n.control
  the required number of study subject in the control group.

  power
  the specified or calculated study power.

Note
Consider a clinical trial comparing two groups, a standard treatment (s) and a new treatment (n). In each group, a proportion of subjects respond to the treatment: $P_s$ and $P_n$.

With a superiority trial we specify the maximum acceptable difference between $P_n$ and $P_s$ as $\delta$. The null hypothesis is $H_0 : P_n - P_s \leq \delta$ and the alternative hypothesis is $H_1 : P_n - P_s > \delta$.

An equivalence trial is used if want to prove that two treatments produce the same clinical outcomes. With an equivalence trial, we specify the maximum acceptable difference between $P_n$ and $P_s$ as $\delta$. The null hypothesis is $H_0 : |P_s - P_n| \geq \delta$ and the alternative hypothesis is $H_1 : |P_s - P_n| < \delta$. In bioequivalence trials, a 90% confidence interval is often used. The value of the maximum acceptable difference $\delta$ is chosen so that a patient will not detect any change in effect when replacing the standard treatment with the new treatment.

With a non-inferiority trial, we specify the maximum acceptable difference between $P_n$ and $P_s$ as $\delta$. The null hypothesis is $H_0 : P_s - P_n \geq \delta$ and the alternative hypothesis is $H_1 : P_s - P_n < \delta$. The aim of a non-inferiority trial is show that a new treatment is not (much) inferior to a standard treatment. Showing non-inferiority can be of interest because: (a) it is often not ethically possible to do a placebo-controlled trial, (b) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints, but is safer, (c) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints, but is cheaper to produce or easier to administer, (d) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints in clinical trial, but compliance will be better outside the clinical trial and hence efficacy better outside the trial.
For a summary of the key features of superiority, equivalence and non-inferiority trials, refer to the documentation for \texttt{epi.ssequb}.

When calculating the power of a study, note that the variable \textit{n} refers to the total study size (that is, the number of subjects in the treatment group plus the number in the control group).

References


Examples

```r
## EXAMPLE 1 (from Machin, Campbell, Tan and Tan 2009 p. 113):
## It is anticipated that patients on a particular drug have a mean diastolic
## blood pressure of 96 mm Hg, against 94 mm Hg on an alternative. It is also
## anticipated that the standard deviation of diastolic BP is approximately
## 8 mm Hg. If one wishes to confirm that the difference is likely to be less
## than 5 mm Hg, that is, one wishes to show equivalence, how many patients
## are need to be enrolled in the trial? Assume 80% power and
## 95% significance.
epi.ssequc(treat = 94, control = 96, sd = 8, delta = 5, n = NA, 
r = 1, power = 0.80, nfractional = FALSE, alpha = 0.05)
## A total of 244 subjects need to be enrolled in the trial, 122 in the
## treatment group and 122 in the control group.

## EXAMPLE 2 (from Chow S, Shao J, Wang H 2008, p. 64):
## A pharmaceutical company is interested in conducting a clinical trial
## to compare two cholesterol lowering agents for treatment of patients with
## congestive heart disease using a parallel design. The primary efficacy
## parameter is the LDL. In what follows, we will consider the situation
## where the intended trial is for testing equivalence of mean responses
## in LDL. Assume that 80% power is required at a 5% level of significance.
## In this example, we assume a 5 unit (i.e. delta = 5) change of LDL is
## considered of clinically meaningful difference. Assume the standard
```
## Sample size for a non-inferiority trial, binary outcome

**Description**

Sample size for a non-inferiority trial, binary outcome.

**Usage**

```r
epi.ssninf(treat, control, delta, n, r = 1, power, nfractional = FALSE, alpha)
```

**Arguments**

- `treat`: the expected proportion of successes in the treatment group.
- `control`: the expected proportion of successes in the control group.
- `delta`: the equivalence limit, expressed as the change in the outcome of interest that represents a clinically meaningful difference.
- `n`: scalar, the total number of study subjects in the trial.
- `r`: scalar, the number in the treatment group divided by the number in the control group.
- `power`: scalar, the required study power.
- `nfractional`: logical, return fractional sample size.
- `alpha`: scalar, defining the desired alpha level.
Value

A list containing the following:

- **n.total**: the total number of study subjects required.
- **n.treat**: the required number of study subject in the treatment group.
- **n.control**: the required number of study subject in the control group.
- **power**: the specified or calculated study power.

Note

Consider a clinical trial comparing two groups, a standard treatment (s) and a new treatment (n). In each group, a proportion of subjects respond to the treatment: $P_s$ and $P_n$.

With a superiority trial we specify the maximum acceptable difference between $P_n$ and $P_s$ as $\delta$. The null hypothesis is $H_0: P_n - P_s \leq \delta$ and the alternative hypothesis is $H_1: P_n - P_s > \delta$.

An equivalence trial is used if want to prove that two treatments produce the same clinical outcomes. With an equivalence trial, we specify the maximum acceptable difference between $P_n$ and $P_s$ as $\delta$. The null hypothesis is $H_0: |P_s - P_n| \geq \delta$ and the alternative hypothesis is $H_1: |P_s - P_n| < \delta$. In bioequivalence trials, a 90% confidence interval is often used. The value of the maximum acceptable difference $\delta$ is chosen so that a patient will not detect any change in effect when replacing the standard treatment with the new treatment.

With a non-inferiority trial, we specify the maximum acceptable difference between $P_n$ and $P_s$ as $\delta$. The null hypothesis is $H_0: P_s - P_n \geq \delta$ and the alternative hypothesis is $H_1: P_s - P_n < \delta$. The aim of a non-inferiority trial is show that a new treatment is not (much) inferior to a standard treatment. Showing non-inferiority can be of interest because: (a) it is often not ethically possible to do a placebo-controlled trial, (b) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints, but is safer, (c) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints, but is cheaper to produce or easier to administer, (d) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints in clinical trial, but compliance will be better outside the clinical trial and hence efficacy better outside the trial.

For a summary of the key features of superiority, equivalence and non-inferiority trials, refer to the documentation for **epi.ssequb**.

When calculating the power of a study, note that the variable $n$ refers to the total study size (that is, the number of subjects in the treatment group plus the number in the control group).

References


Examples

## A pharmaceutical company would like to conduct a clinical trial to
## compare the efficacy of two antimicrobial agents when administered orally
## to patients with skin infections. Assume the true mean cure rate of the
## treatment is 0.85 and the true mean cure rate of the control is 0.65.
## We consider a difference of less than 0.10 in cure rate to be of no
## clinical importance (i.e. delta = -0.10).
## Assuming a one-sided test size of 5% and a power of 80% how many
## subjects should be included in the trial?

epi.ssninfb(treat = 0.85, control = 0.65, delta = -0.10, n = NA, r = 1,
            power = 0.80, nfractional = FALSE, alpha = 0.05)

## A total of 50 subjects need to be enrolled in the trial, 25 in the
## treatment group and 25 in the control group.

## EXAMPLE 1 (cont.):
## Suppose only 40 subjects were enrolled in the trial, 20 in the treatment
## group and 20 in the control group. What is the estimated study power?

epi.ssninfb(treat = 0.85, control = 0.65, delta = -0.10, n = 40, r = 1,
            power = NA, nfractional = FALSE, alpha = 0.05)

## With only 40 subjects the estimated study power is 0.73.

---

epi.ssninfc  
Sample size for a non-inferiority trial, continuous outcome

### Description

Sample size for a non-inferiority trial, continuous outcome.

### Usage

epi.ssninfc(treat, control, sd, delta, n, r = 1, power, nfractional = FALSE, alpha)
Arguments

- **treat**: the expected mean of the outcome of interest in the treatment group.
- **control**: the expected mean of the outcome of interest in the control group.
- **sd**: the expected population standard deviation of the outcome of interest.
- **delta**: the equivalence limit, expressed as the change in the outcome of interest that represents a clinically meaningful difference.
- **n**: scalar, the total number of study subjects in the trial.
- **r**: scalar, the number in the treatment group divided by the number in the control group.
- **power**: scalar, the required study power.
- **nfractional**: logical, return fractional sample size.
- **alpha**: scalar, defining the desired alpha level.

Value

A list containing the following:

- **n.total**: the total number of study subjects required.
- **n.treat**: the required number of study subject in the treatment group.
- **n.control**: the required number of study subject in the control group.
- **power**: the specified or calculated study power.

Note

Consider a clinical trial comparing two groups, a standard treatment (s) and a new treatment (n). In each group, a proportion of subjects respond to the treatment: \( P_s \) and \( P_n \).

With a superiority trial we specify the maximum acceptable difference between \( P_n \) and \( P_s \) as \( \delta \). The null hypothesis is \( H_0 : P_n - P_s \leq \delta \) and the alternative hypothesis is \( H_1 : P_n - P_s > \delta \).

An equivalence trial is used if want to prove that two treatments produce the same clinical outcomes. With an equivalence trial, we specify the maximum acceptable difference between \( P_n \) and \( P_s \) as \( \delta \). The null hypothesis is \( H_0 : |P_s - P_n| \geq \delta \) and the alternative hypothesis is \( H_1 : |P_s - P_n| < \delta \). In bioequivalence trials, a 90% confidence interval is often used. The value of the maximum acceptable difference \( \delta \) is chosen so that a patient will not detect any change in effect when replacing the standard treatment with the new treatment.

With a non-inferiority trial, we specify the maximum acceptable difference between \( P_n \) and \( P_s \) as \( \delta \). The null hypothesis is \( H_0 : P_s - P_n \geq \delta \) and the alternative hypothesis is \( H_1 : P_s - P_n < \delta \). The aim of a non-inferiority trial is show that a new treatment is not (much) inferior to a standard treatment. Showing non-inferiority can be of interest because: (a) it is often not ethically possible to do a placebo-controlled trial, (b) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints, but is safer, (c) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints, but is cheaper to produce or easier to administer, (d) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints in clinical trial, but compliance will be better outside the clinical trial and hence efficacy better outside the trial.
For a summary of the key features of superiority, equivalence and non-inferiority trials, refer to the documentation for epi.ssequb.

When calculating the power of a study, note that the variable n refers to the total study size (that is, the number of subjects in the treatment group plus the number in the control group).

References


Examples

## A pharmaceutical company is interested in conducting a clinical trial
to compare two cholesterol lowering agents for treatment of patients with
congestive heart disease using a parallel design. The primary efficacy
parameter is the LDL. In what follows, we will consider the situation
where the intended trial is for testing non-inferiority of mean responses
in LDL. Assume that 80% power is required at a 5% level of significance.

In this example, we assume a 5% (i.e. \( \delta = 0.05 \)) change in LDL is
a clinically meaningful difference. Assume the standard deviation of
LDL is 0.10 and the LDL concentration in the treatment group is 0.20
units and the LDL concentration in the control group is 0.20 units.

\[
\text{epi.ssninf}(treat = 0.20, \text{control} = 0.20, sd = 0.10, \delta = 0.05, n = \text{NA}, r = 1, \text{power} = 0.80, \text{nfractional} = \text{FALSE}, \alpha = 0.05)
\]

A total of 100 subjects need to be enrolled in the trial, 50 in the
treatment group and 50 in the control group.
Sample size to estimate a binary outcome using simple random sampling.

Usage

epi.sssimpleestb(N = 1E+06, Py, epsilon, error = "relative", se, sp, nfractional = FALSE, conf.level = 0.95)

Arguments

N scalar integer, the total number of individual listing units in the population.
Py scalar number, an estimate of the population proportion to be estimated.
epsilon scalar number, the maximum difference between the estimate and the unknown population value expressed in absolute or relative terms.
error character string. Options are absolute for absolute error and relative for relative error.
se the diagnostic sensitivity of the method used to detect positive outcomes (0 - 1).
sp the diagnostic specificity of the method used to detect positive outcomes (0 - 1).
fractional logical, return fractional sample size.
conf.level scalar number, the level of confidence in the computed result.

Value

Returns an integer defining the required sample size.

Note

The sample size calculation method implemented in this function follows the approach described by Humphry et al. (2004) accounting for imperfect diagnostic sensitivity and specificity.

If epsilon.r equals the relative error the sample estimate should not differ in absolute value from the true unknown population parameter d by more than epsilon.r * d.

References


Examples

## EXAMPLE 1:
## We want to estimate the seroprevalence of Brucella abortus in a population
## of cattle. An estimate of the unknown prevalence of B. abortus in this
## population is 0.15. We would like to be 95% certain that our estimate is
## within 20% of the true proportion of the population seropositive to
## B. abortus. Calculate the required sample size assuming use of a test
## with perfect diagnostic sensitivity and specificity.

def n.crude <- epi.sssimpleestb(N = 1E+06, Py = 0.15, epsilon = 0.20,
                           error = "relative", se = 1.00, sp = 1.00, nfractional = FALSE,
                           conf.level = 0.95)
n.crude
## A total of 545 cattle need to be sampled to meet the requirements of the
## survey.

## EXAMPLE 1 (continued):
## Why don't I get the same results as other sample size calculators? The
## most likely reason is misspecification of epsilon. Other sample size
## calculators (e.g. OpenEpi) require you to enter the absolute
## error (as opposed to relative error). For the example above the absolute
## error is 0.20 * 0.15 = 0.03. Re-run epi.simpleestb:

def n.crude <- epi.sssimpleestb(N = 1E+06, Py = 0.15, epsilon = 0.03,
                           error = "absolute", se = 0.94, sp = 0.88, nfractional = FALSE,
                           conf.level = 0.95)
n.crude
## A total of 545 cattle need to be sampled to meet the requirements of the
## survey.

## EXAMPLE 1 (continued):
## The OIE recommends that the compliment fixation test (CFT) is used for
## bovine brucellosis prevalence estimation. Assume the diagnostic sensitivity
## and specificity of the bovine brucellosis CFT to be used is 0.94 and 0.88
## respectively (Getachew et al. 2016). Re-calculate the required sample size
## accounting for imperfect diagnostic test performance.

def n.crude <- epi.sssimpleestb(N = 1E+06, Py = 0.15, epsilon = 0.20,
                           error = "relative", se = 0.94, sp = 0.88, nfractional = FALSE,
Sample size to estimate a continuous outcome using simple random sampling

Description

Sample size to estimate a continuous outcome using simple random sampling.

Usage

epi.sssimpleestc(N = 1E+06, xbar, sigma, epsilon, error = "relative", nfrational = FALSE, conf.level = 0.95)
Arguments

N  scalar integer, representing the total number of individual listing units in the population.

xbar  scalar number, the expected mean of the continuous variable to be estimated.

sigma  scalar number, the expected standard deviation of the continuous variable to be estimated.

epsilon  scalar number, the maximum difference between the estimate and the unknown population value expressed in absolute or relative terms.

error  character string. Options are absolute for absolute error and relative for relative error.

nfractional  logical, return fractional sample size.

cnf.level  scalar number, the level of confidence in the computed result.

Value

Returns an integer defining the required sample size.

Note

If epsilon.r equals the relative error the sample estimate should not differ in absolute value from the true unknown population parameter \( \theta \) by more than epsilon.r * \( \theta \).

References


Examples

```r
## EXAMPLE 1:
## A city contains 20 neighbourhood health clinics and it is desired to take a
## sample of clinics to estimate the total number of persons from all these
## clinics who have been given, during the past 12 month period, prescriptions
## for a recently approved antidepressant. If we assume that the average number
## of people seen at these clinics is 1500 per year with the standard deviation
## equal to 300, and that approximately 5% of patients (regardless of clinic)
## are given this drug, how many clinics need to be sampled to yield an estimate
## that is within 20% of the true population value?

pmean <- 1500 * 0.05; psigma <- (300 * 0.05)
epi.ssimpleestc(N = 20, xbar = pmean, sigma = psigma, epsilon = 0.20,
error = "relative", nfractional = FALSE, conf.level = 0.95)
```
## Four clinics need to be sampled to meet the requirements of the survey.

## EXAMPLE 2:
## We want to estimate the mean body weight of deer on a farm. There are 278 animals present. We anticipate the mean body weight to be around 200 kg and the standard deviation of body weight to be 30 kg. We would like to be 95% certain that our estimate is within 10 kg of the true mean. How many deer should be sampled?

```r
epi.sssimpleestc(N = 278, xbar = 200, sigma = 30, epsilon = 10, error = "absolute", nfractional = FALSE, conf.level = 0.95)
```

## A total of 31 deer need to be sampled to meet the requirements of the survey.

---

**epi.ssstrataestb**  
*Sample size to estimate a binary outcome using stratified random sampling*

### Description
Sample size to estimate a binary outcome using stratified random sampling.

### Usage
```r
epi.ssstrataestb(strata.n, strata.Py, epsilon, error = "relative", nfractional = FALSE, conf.level = 0.95)
```

### Arguments
- `strata.n`: vector of integers, the number of individual listing units in each strata.
- `strata.Py`: vector of numbers, the expected proportion of individual listing units with the outcome of interest for each strata.
- `epsilon`: scalar number, the maximum difference between the estimate and the unknown population value expressed in absolute or relative terms.
- `error`: character string. Options are absolute for absolute error and relative for relative error.
- `nfractional`: logical, return fractional sample size.
- `conf.level`: scalar number, the level of confidence in the computed result.

### Value
A list containing the following:
- `strata.sample`: the estimated sample size for each strata.
- `strata.total`: the estimated total size.
strata.stats  mean the mean across all strata, sigma.bx the among-strata variance, sigma.wx the within-strata variance, and sigma.x the among-strata variance plus the within-strata variance, rel.var the within-strata variance divided by the square of the mean, and gamma the ratio of among-strata variance to within-strata variance.

Author(s)
Mark Stevenson (Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Australia).
Javier Sanchez (Atlantic Veterinary College, University of Prince Edward Island, Charlottetown Prince Edward Island, C1A 4P3, Canada).

References

Examples
## EXAMPLE 1:
## Dairies are to be sampled to determine the proportion of herd managers
## using foot bathes. Herds are stratified according to size (small, medium,
## and large). The number of herds in each strata are 1500, 2500, and 4000
## (respectively). A review of the literature indicates that use of foot bathes
## on farms is in the order of 0.50, with the probability of usage increasing
## as herds get larger. How many dairies should be sampled?
strata.n <- c(1500, 2500, 4000)
strata.Py <- c(0.50, 0.60, 0.70)
epi.ssstrataestb(strata.n, strata.Py, epsilon = 0.20, error = "relative",
nfractional = FALSE, conf.level = 0.95)

## A total of 55 herds should be sampled: 11 small, 18 medium, and 28 large.

epi.ssstrataestc  Sample size to estimate a continuous outcome using a stratified random sampling design

Description
Sample size to estimate a continuous outcome using a stratified random sampling design.

Usage
epi.ssstrataestc(strata.n, strata.xbar, strata.sigma, epsilon,
error = "relative", nfractional = FALSE, conf.level = 0.95)
Arguments

- **strata.n** vector of integers, defining the number of individual listing units in each strata.
- **strata.xbar** vector of numbers, defining the expected means of the continuous variable to be estimated for each strata.
- **strata.sigma** vector of numbers, defining the expected standard deviation of the continuous variable to be estimated for each strata.
- **epsilon** scalar number, the maximum difference between the estimate and the unknown population value expressed in absolute or relative terms.
- **error** character string. Options are absolute for absolute error and relative for relative error.
- **nfractional** logical, return fractional sample size.
- **conf.level** scalar number, the level of confidence in the computed result.

Value

A list containing the following:

- **strata.sample** the estimated sample size for each strata.
- **strata.total** the estimated total size.
- **strata.stats** mean the mean across all strata, sigma.bx the among-strata variance, sigma.wx the within-strata variance, and sigma.x the among-strata variance plus the within-strata variance, rel.var the within-strata variance divided by the square of the mean, and gamma the ratio of among-strata variance to within-strata variance.

Author(s)

Mark Stevenson (Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Australia).

Javier Sanchez (Atlantic Veterinary College, University of Prince Edward Island, Charlottetown Prince Edward Island, C1A 4P3, Canada).

References


Examples

## EXAMPLE 1 (from Levy and Lemeshow 1999, page 176 -- 178):

```r
## We plan to take a sample of the members of a health maintenance
## organisation (HMO) for purposes of estimating the average number
## of hospital episodes per person per year. The sample will be selected
## from membership lists according to age (under 45 years, 45 -- 64 years,
## 65 years and over). The number of members in each strata are 600, 500,
## and 400 (respectively). Previous data estimates the mean number of
## hospital episodes per year for each strata as 0.164, 0.166, and 0.236
## (respectively). The variance of these estimates are 0.245, 0.296, and
```
## 0.436 (respectively). How many from each strata should be sampled to be
## 95% that the sample estimate of hospital episodes is within 20% of the
## true value?

```r
c(600, 500, 400)
c(0.164, 0.166, 0.236)
c(0.245, 0.296, 0.436))
epi.ssstrataestc(c(0.164, 0.166, 0.236),
    epsilon = 0.20,
    error = "relative", nfractional = FALSE, conf.level = 0.95)
```

## The number allocated to the under 45 years, 45 -- 64 years, and 65 years
## and over stratums should be 224, 187, and 150 (a total of 561). These
## results differ from the worked example provided in Levy and Lemeshow where
## certainty is set to approximately 99%.

---

epi.sssupb

### Sample size for a parallel superiority trial, binary outcome

#### Description

Sample size for a parallel superiority trial, binary outcome.

#### Usage

```r
epi.sssupb(treat, control, delta, n, r = 1, power, nfractional = FALSE, alpha)
```

#### Arguments

- **treat**: the expected proportion of successes in the treatment group.
- **control**: the expected proportion of successes in the control group.
- **delta**: the equivalence limit, expressed as the change in the outcome of interest that represents a clinically meaningful difference.
- **n**: scalar, the total number of study subjects in the trial.
- **r**: scalar, the number in the treatment group divided by the number in the control group.
- **power**: scalar, the required study power.
- **nfractional**: logical, return fractional sample size.
- **alpha**: scalar, defining the desired alpha level.

#### Value

A list containing the following:

- **n.total**: the total number of study subjects required.
- **n.treat**: the required number of study subject in the treatment group.
- **n.control**: the required number of study subject in the control group.
- **power**: the specified or calculated study power.
Note

Consider a clinical trial comparing two groups, a standard treatment (s) and a new treatment (n). In each group, a proportion of subjects respond to the treatment: $P_s$ and $P_n$.

With a superiority trial we specify the maximum acceptable difference between $P_n$ and $P_s$ as $\delta$. The null hypothesis is $H_0: P_n - P_s \leq \delta$ and the alternative hypothesis is $H_1: P_n - P_s > \delta$.

An equivalence trial is used if we want to prove that two treatments produce the same clinical outcomes. With an equivalence trial, we specify the maximum acceptable difference between $P_n$ and $P_s$ as $\delta$. The null hypothesis is $H_0: |P_s - P_n| \geq \delta$ and the alternative hypothesis is $H_1: |P_s - P_n| < \delta$. In bioequivalence trials, a 90% confidence interval is often used. The value of the maximum acceptable difference $\delta$ is chosen so that a patient will not detect any change in effect when replacing the standard treatment with the new treatment.

With a non-inferiority trial, we specify the maximum acceptable difference between $P_n$ and $P_s$ as $\delta$. The null hypothesis is $H_0: P_s - P_n \geq \delta$ and the alternative hypothesis is $H_1: P_s - P_n < \delta$. The aim of a non-inferiority trial is to show that a new treatment is not (much) inferior to a standard treatment. Showing non-inferiority can be of interest because: (a) it is often not ethically possible to do a placebo-controlled trial, (b) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints, but is safer, (c) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints, but is cheaper to produce or easier to administer, (d) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints in clinical trial, but compliance will be better outside the clinical trial and hence efficacy better outside the trial.

For a summary of the key features of superiority, equivalence and non-inferiority trials, refer to the documentation for epi.ssequb.

When calculating the power of a study, note that the variable $n$ refers to the total study size (that is, the number of subjects in the treatment group plus the number in the control group).

References


Examples


## Suppose that a pharmaceutical company is interested in conducting a clinical trial to compare the efficacy of two antimicrobial agents when administered orally once daily in the treatment of patients with skin infections. In what follows, we consider the situation where the intended trial is for testing superiority of the test drug over the active control drug. For this purpose, the following assumptions are made. First, sample size calculation will be performed for achieving 80% power at the 5% level of significance.

## Assume the true mean cure rates of the treatment agents and the active
## control are 85% and 65%, respectively. Assume the superiority
## margin is 5%.

epi.sssupb(treat = 0.85, control = 0.65, delta = 0.05, n = NA,
           r = 1, power = 0.80, nfractional = FALSE, alpha = 0.05)

## A total of 196 subjects need to be enrolled in the trial, 98 in the
## treatment group and 98 in the control group.

epi.sssupc  

---

### Sample size for a parallel superiority trial, continuous outcome

**Description**

Sample size for a parallel superiority trial, continuous outcome.

**Usage**

epi.sssupc(treat, control, sd, delta, n, r = 1, power, nfractional = FALSE,
           alpha)

**Arguments**

- **treat**: the expected mean of the outcome of interest in the treatment group.
- **control**: the expected mean of the outcome of interest in the control group.
- **sd**: the expected population standard deviation of the outcome of interest.
- **delta**: the equivalence limit, expressed as the change in the outcome of interest that represents a clinically meaningful difference.
- **n**: scalar, the total number of study subjects in the trial.
- **r**: scalar, the number in the treatment group divided by the number in the control group.
- **power**: scalar, the required study power.
- **nfractional**: logical, return fractional sample size.
- **alpha**: scalar, defining the desired alpha level.

**Value**

A list containing the following:

- **n.total**: the total number of study subjects required.
- **n.treat**: the required number of study subject in the treatment group.
- **n.control**: the required number of study subject in the control group.
- **power**: the specified or calculated study power.
Note

Consider a clinical trial comparing two groups, a standard treatment (s) and a new treatment (n). In each group, a proportion of subjects respond to the treatment: \( P_s \) and \( P_n \).

With a superiority trial we specify the maximum acceptable difference between \( P_n \) and \( P_s \) as \( \delta \). The null hypothesis is \( H_0 : P_n - P_s \leq \delta \) and the alternative hypothesis is \( H_1 : P_n - P_s > \delta \).

An equivalence trial is used if we want to prove that two treatments produce the same clinical outcomes. With an equivalence trial, we specify the maximum acceptable difference between \( P_n \) and \( P_s \) as \( \delta \). The null hypothesis is \( H_0 : |P_s - P_n| \geq \delta \) and the alternative hypothesis is \( H_1 : |P_s - P_n| < \delta \). In bioequivalence trials, a 90% confidence interval is often used. The value of the maximum acceptable difference \( \delta \) is chosen so that a patient will not detect any change in effect when replacing the standard treatment with the new treatment.

With a non-inferiority trial, we specify the maximum acceptable difference between \( P_n \) and \( P_s \) as \( \delta \). The null hypothesis is \( H_0 : P_s - P_n \geq \delta \) and the alternative hypothesis is \( H_1 : P_s - P_n < \delta \). The aim of a non-inferiority trial is to show that a new treatment is not (much) inferior to a standard treatment. Showing non-inferiority can be of interest because: (a) it is often not ethically possible to do a placebo-controlled trial, (b) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints, but is safer, (c) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints, but is cheaper to produce or easier to administer, (d) the new treatment is not expected to be better than the standard treatment on primary efficacy endpoints in clinical trial, but compliance will be better outside the clinical trial and hence efficacy better outside the trial.

For a summary of the key features of superiority, equivalence and non-inferiority trials, refer to the documentation for epi.ssequb.

When calculating the power of a study, note that the variable \( n \) refers to the total study size (that is, the number of subjects in the treatment group plus the number in the control group).

References


Examples

```r
## EXAMPLE 1:
## A pharmaceutical company is interested in conducting a clinical trial
## to compare two cholesterol lowering agents for treatment of patients with
## congestive heart disease (CHD) using a parallel design. The primary
## efficacy parameter is the concentration of high density lipoproteins.
## (HDL). We consider the situation where the intended trial is to test
## superiority of the test drug over the active control agent. Sample
## size calculations are to be calculated to achieve 80% power at the
## 5% level of significance.
##
## In this example, we assume that if treatment results in a 5 unit
```
(i.e. delta = 5) increase in HDL it is declared to be superior to the active control. Assume the standard deviation of HDL is 10 units and the HDL concentration in the treatment group is 20 units and the HDL concentration in the control group is 20 units.

```r
epi.sssupc(treat = 20, control = 20, sd = 10, delta = 5, n = NA, r = 1, power = 0.80, nfractional = FALSE, alpha = 0.05)
```

A total of 100 subjects need to be enrolled in the trial, 50 in the treatment group and 50 in the control group.

---

### Sample size, power or minimum detectable prevalence ratio or odds ratio for a cross-sectional study

**Description**

Sample size, power or minimum detectable prevalence ratio or odds ratio for a cross-sectional study.

**Usage**

```r
epi.ssxsectn(pdexp1, pdexp0, pexp = NA, n = NA, power = 0.80, r = 1, N, design = 1, sided.test = 2, finite.correction = FALSE, nfractional = FALSE, conf.level = 0.95)
```

**Arguments**

- `pdexp1`: the expected prevalence of the outcome in the exposed group (0 to 1).
- `pdexp0`: the expected prevalence of the outcome in the non-exposed group (0 to 1).
- `pexp`: the expected prevalence of exposure to the hypothesised risk factor in the population (0 to 1).
- `n`: scalar, defining the total number of subjects in the study (i.e. the number in both the exposed and unexposed groups).
- `power`: scalar, the required study power.
- `r`: scalar, the number in the exposed group divided by the number in the unexposed group.
- `N`: scalar, the estimated number of individuals in the population.
- `design`: scalar, the estimated design effect.
- `sided.test`: use a one- or two-sided test? Use a two-sided test if you wish to evaluate whether or not the outcome incidence risk in the exposed group is greater than or less than the outcome incidence risk in the unexposed group. Use a one-sided test to evaluate whether or not the outcome incidence risk in the exposed group is greater than the outcome incidence risk in the unexposed group.
- `finite.correction`: logical, apply a finite correction factor?
- `nfractional`: logical, return fractional sample size.
- `conf.level`: scalar, defining the level of confidence in the computed result.
Details

The methodology in this function follows the approach described in Chapter 8 of Woodward (2005), pp. 381 - 426.

Value

A list containing the following:

- `n.total`: the total number of subjects required for the specified level of confidence and power, respecting the requirement for \( r \) times as many individuals in the exposed (treatment) group compared with the non-exposed (control) group.
- `n.exp1`: the total number of subjects in the exposed (treatment) group for the specified level of confidence and power, respecting the requirement for \( r \) times as many individuals in the exposed (treatment) group compared with the non-exposed (control) group.
- `n.exp0`: the total number of subjects in the non-exposed (control) group for the specified level of confidence and power, respecting the requirement for \( r \) times as many individuals in the exposed (treatment) group compared with the non-exposed (control) group.
- `power`: the power of the study given the number of study subjects, the expected effect size and level of confidence.
- `pr`: the prevalence of the outcome in the exposed group divided by the prevalence of the outcome in the unexposed group (the prevalence ratio).
- `or`: the odds of the outcome in the exposed group divided by the odds of the outcome in the unexposed group (the odds ratio).

Note

The power of a study is its ability to demonstrate the presence of an association, given that an association actually exists.

Values need to be entered for `pdxp0`, `pexp`, `n`, and `power` to return a value for the prevalence ratio `pr` and odds ratio `or`. In this situation, the lower value of `pr` represents the maximum detectable prevalence ratio that is less than 1; the upper value of `pr` represents the minimum detectable prevalence ratio greater than 1. A value for `pexp` doesn’t need to be entered if you want to calculate sample size or study power.

When calculating study power or minimum detectable prevalence risk ratio when `finite.correction` = TRUE the function takes the values of `n` and `N` entered by the user and back-calculates a value of `n` assuming an infinite population. Values for `power`, `pr` and `or` are then returned, assuming the back-calculated value of `n` is equivalent to the value of `n` entered by the user.

See the documentation for `epi.sscohortc` for an example using the design facility implemented in this function.

References


Examples

```r
## EXAMPLE 1:
## A cross-sectional study is to be carried out to quantify the association
## between farm management type (intensive, extensive) and evidence of
## Q fever in dairy goat herds. The investigators would like to be 0.80 sure
## of being able to detect when the risk ratio of Q fever is 2.0 for
## intensively managed herds, using a 0.05 significance test. Previous evidence
## suggests that the prevalence of Q fever in extensively managed dairy goat
## herds is 5 per 100 herds at risk and the prevalence of intensively managed
## herds in the population (the prevalence of exposure) is around 0.20.
##
## Assuming equal numbers of intensively managed and extensively managed
## herds will be sampled, how many herds need to be enrolled into the study?
## You estimate that there are around 60 dairy goat herds in your study area.

pdexp1 = 2.0 * (5 / 100); pdexp0 = 5 / 100
epi.ssxsectn(pdexp1 = pdexp1, pdexp0 = pdexp0, pexp = 0.20, n = NA,
power = 0.80, r = 1, N = 60, design = 1, sided.test = 2,
finite.correction = TRUE, nfractional = FALSE, conf.level = 0.95)

## A total of 58 of the 60 herds need to be enrolled into the study
## (29 intensively managed and 29 extensively managed herds).

## EXAMPLE 2:
## Say, for example, we're only able to enrol 45 herds into the study
## described above. What is the minimum and maximum detectable prevalence
## ratio and minimum and maximum detectable odds ratio?

epi.ssxsectn(pdexp1 = NA, pdexp0 = pdexp0, pexp = 0.20, n = 45,
power = 0.80, r = 1, N = 60, design = 1, sided.test = 2,
finite.correction = TRUE, nfractional = FALSE, conf.level = 0.95)

## The minimum detectable prevalence ratio >1 is 3.64. The maximum detectable
## prevalence ratio <1 is 0.

## The minimum detectable odds ratio >1 is 4.65. The maximum detectable
## odds ratio <1 is 0.
```
Description

Computes true and apparent prevalence, sensitivity, specificity, positive and negative predictive values and positive and negative likelihood ratios from count data provided in a 2 by 2 table.

Usage

epi.tests(dat, conf.level = 0.95)

## S3 method for class 'epi.tests'
print(x, ...)

## S3 method for class 'epi.tests'
summary(object, ...)

Arguments

dat a vector of length four, an object of class table or an object of class grouped_df from package dplyr containing the individual cell frequencies (see below).

conf.level magnitude of the returned confidence interval. Must be a single number between 0 and 1.

x, object an object of class epi.tests.

... Ignored.

Details

Exact binomial confidence limits are calculated for test sensitivity, specificity, and positive and negative predictive value (see Collett 1999 for details).

Confidence intervals for positive and negative likelihood ratios are based on formulae provided by Simel et al. (1991).

Diagnostic accuracy is defined as the proportion of all tests that give a correct result. Diagnostic odds ratio is defined as how much more likely will the test make a correct diagnosis than an incorrect diagnosis in patients with the disease (Scott et al. 2008). The number needed to diagnose is defined as the number of patients that need to be tested to give one correct positive test. Youden’s index is the difference between the true positive rate and the false positive rate. Youden’s index ranges from -1 to +1 with values closer to 1 if both sensitivity and specificity are high (i.e. close to 1).

Value

An object of class epi.tests containing the following:

ap apparent prevalence.

tp true prevalence.

se test sensitivity.

sp test specificity.

diag.ac diagnostic accuracy.

diag.or diagnostic odds ratio.
ndx number needed to diagnose.
youden Youden’s index.
\( pv. pos \) positive predictive value.
\( pv. neg \) negative predictive value.
\( lr. pos \) likelihood ratio of a positive test.
\( lr. neg \) likelihood ratio of a negative test.
\( p. rout \) the proportion of subjects with the outcome ruled out.
\( p. rin \) the proportion of subjects with the outcome ruled in.
\( p. fpos \) of all the subjects that are truly outcome negative, the proportion that are incorrectly classified as positive (the proportion of false positives).
\( p. fneg \) of all the subjects that are truly outcome positive, the proportion that are incorrectly classified as negative (the proportion of false negative).

**Note**

<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test +</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Test -</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

**Author(s)**

Mark Stevenson (Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Australia). Charles Reynard (School of Medical Sciences, The University of Manchester, United Kingdom).

**References**


Examples

## EXAMPLE 1 (from Scott et al. 2008, Table 1):
## A new diagnostic test was trialled on 1586 patients. Of 744 patients that
## were disease positive, 670 were test positive. Of 842 patients that were
## disease negative, 640 were test negative. What is the likelihood ratio of
## a positive test? What is the number needed to diagnose?

dat.v01 <- c(670,202,74,640)
rval.tes01 <- epi.tests(dat.v01, conf.level = 0.95)
print(rval.tes01)

## Test sensitivity is 0.90 (95% CI 0.88 to 0.92). Test specificity is
## 0.76 (95% CI 0.73 to 0.79). The likelihood ratio of a positive test
## is 3.75 (95% CI 3.32 to 4.24).

## What is the number needed to diagnose?
names(rval.tes01$detail)
rval.tes01$detail$nndx

## The number needed to diagnose is 1.51 (95% CI 1.41 to 1.65). Around 15
## persons need to be tested to return 10 positive tests.

## EXAMPLE 2:
## Same as Example 1 but showing how a 2 by 2 contingency table can be prepared
## using tidyverse:

## Not run:
library(tidyverse)

generate a data set listing test results and true disease status:
dis <- c(rep(1, times = 744), rep(0, times = 842))
tes <- c(rep(1, times = 670), rep(0, times = 74),
     rep(1, times = 202), rep(0, times = 640))
dat.df02 <- data.frame(dis, tes)
tmp.df02 <- dat.df02 %>%
  mutate(dis = factor(dis, levels = c(1,0), labels = c("Dis+","Dis-"))) %>%
  mutate(tes = factor(tes, levels = c(1,0), labels = c("Test+","Test-"))) %>%
  group_by(tes, dis) %>%
  summarise(n = n())
tmp.df02

## View the data in conventional 2 by 2 table format:
pivot_wider(tmp.df02, id_cols = c(tes), names_from = dis, values_from = n)

rval.tes02 <- epi.tests(tmp.df02, conf.level = 0.95)
print(rval.tes02)
## Test sensitivity is 0.90 (95% CI 0.88 to 0.92). Test specificity is
## 0.76 (95% CI 0.73 to 0.79). The likelihood ratio of a positive test
## is 3.75 (95% CI 3.32 to 4.24).

### EXAMPLE 3:
### A biomarker assay has been developed to identify patients that are at
### high risk of experiencing myocardial infarction. The assay varies on
### a continuous scale, from 0 to 1. Researchers believe that a biomarker
### assay result of greater than or equal to 0.60 renders a patient test
### positive, that is, at elevated risk of experiencing a heart attack
### over the next 12 months.

### Generate data consistent with the information provided above. Assume the
### prevalence of high risk subjects in your population is 0.35:
```r
data.frame(out = rbinom(n = 200, size = 1, prob = 0.35),
            bm = runif(n = 200, min = 0, max = 1))
```

### Classify study subjects as either test positive or test negative
### according to their biomarker test result:
```r
data.frame(out = rbinom(n = 200, size = 1, prob = 0.35),
            bm = runif(n = 200, min = 0, max = 1))
```

### Generate a two-by-two table:
```r
table(out = rbinom(n = 200, size = 1, prob = 0.35),
       bm = runif(n = 200, min = 0, max = 1))
```

### What proportion of subjects are ruled out as being at high risk of
### myocardial infarction?
```r
data.frame(out = rbinom(n = 200, size = 1, prob = 0.35),
            bm = runif(n = 200, min = 0, max = 1))
```

### What proportion of subjects are ruled in as being at high risk of
### myocardial infarction?
```r
data.frame(out = rbinom(n = 200, size = 1, prob = 0.35),
            bm = runif(n = 200, min = 0, max = 1))
```

### What is the proportion of false positive results?
```r
data.frame(out = rbinom(n = 200, size = 1, prob = 0.35),
            bm = runif(n = 200, min = 0, max = 1))
```

### What is the proportion of false negative results?
```r
data.frame(out = rbinom(n = 200, size = 1, prob = 0.35),
            bm = runif(n = 200, min = 0, max = 1))
```
**Description**

Calculates adjusted risk estimates for given relative risk and population proportions. This is an intermediate calculation in the calculation of effective probability of infection for risk-based surveillance activities.

**Usage**

```r
rsu.adjrisk(rr, ppr)
```

**Arguments**

- `rr` vector or matrix, defining the relative risk values for each strata in the population. See details.
- `ppr` vector of length `rr` defining the population proportions in each strata.

**Details**

On some occasions there is interest in calculating adjusted risk values for a series of relative risk estimates drawn from (for example) a probability distribution. In this situation a matrix is passed to argument `rr` with the columns of the matrix corresponding to the number of risk strata and the rows corresponding to the number of iterations for simulation. When data are entered in this format `rsu.adjrisk` returns a matrix of adjusted risk values of the same dimension. See Example 3, below.

**Value**

A vector of adjusted risk values listed in order of `rr`.

**References**


**Examples**

```r
## EXAMPLE 1:
## The relative risk of a given disease in an area of your country is 5
## compared with a known reference 'low risk' area. A recent census shows that
## 10% of the population are resident in the high risk area and 90%
## are resident in the low risk area. .
##
## Calculate the adjusted relative risks for each area.

rsu.adjrisk(rr = c(5,1), ppr = c(0.10,0.90))
```

```r
## The adjusted relative risks for the high and low risk areas are 3.6 and
## 0.7, respectively.

## EXAMPLE 2:
```
## Re-calculate the adjusted relative risks assuming there are 'high',
## 'medium' and 'low' risk areas. The relative risks for the high, medium
## and low risk areas are 5, 3 and 1, respectively. Population proportions for
## each area are 0.10, 0.10 and 0.80, respectively.

rsu.adjrisk(rr = c(5,3,1), ppr = c(0.10,0.10,0.80))

## The adjusted relative risks for the high, medium and low risk areas are
## 3.1, 1.9 and 0.6, respectively.

## EXAMPLE 3:
## Consider now the situation where we are not certain of our relative risk
## estimates for the high, medium and low risk areas described in Example 2
## so we ask a group of experts for their opinion. Minimum, mode and maximum
## relative risk estimates for the high and medium risk areas are defined
## using a PERT distribution. For the high risk area the mode of the
## relative risk is 5 with a minimum of 3 and a maximum of 20. For the medium
## risk area the mode of the relative risk is 3 with a minimum of 2 and a
## maximum of 20. As before, the population proportions for each area are
## 0.10, 0.10 and 0.80, respectively. Take 10 random draws from a PERT
## distribution (using the rpert function in package mc2d) and calculate
## the adjusted relative risks for each draw:

## Not run:
## Set up an empty matrix to collect the simulated relative risk values:
nsims <- 10; nrcat <- 3
rr <- matrix(NA, nrow = nsims, ncol = nrcat)

## Use the mc2d package to take nsims random draws from the PERT distribution:
rr[,1] <- mc2d::rpert(n = nsims, min = 3, mode = 5, max = 20)
rr[,2] <- mc2d::rpert(n = nsims, min = 2, mode = 3, max = 5)

## The low risk area is the reference, so its relative risk values are 1:
rr[,3] <- 1

## Population proportions:
ppr <- c(0.10,0.10,0.80)
rval.df <- rsu.adjrisk(rr, ppr)
summary(rval.df)

## The median adjusted relative risks for the high, medium and low risk area
## are 3.6, 1.6 and 0.5 (respectively). The minimum adjusted relative risks
## are 2.5, 1.3 and 0.39, respectively. The maximum adjusted relative risks
## are 5.5, 2.3 and 0.72, respectively.

## End(Not run)
rsu.dxtest  Sensitivity and specificity of diagnostic tests interpreted in series or parallel

Description

Calculates the sensitivity and specificity of two diagnostic tests interpreted in series or parallel.

Usage

rsu.dxtest(se, sp, interpretation = "series", covar = c(0,0))

Arguments

- se: a vector of length two defining the diagnostic sensitivity of the two tests.
- sp: a vector of length two defining the diagnostic specificity of the two tests.
- interpretation: a character string indicating how the test results should be interpreted. Options are series or parallel.
- covar: a vector of length two defining the covariance between test results for disease positive and disease negative groups. The first element of the vector is the covariance between test results for disease positive subjects. The second element of the vector is the covariance between test results for disease negative subjects. Use covar = c(0,0) (the default) if the tests are known to be independent.

Value

A list comprised of two elements:

- independent: a data frame listing sensitivity se and specificity sp assuming the tests are independent.
- dependent: a data frame listing sensitivity se and specificity sp calculated using the values of covar, as entered by the user.

If covar = c(0,0) data frames independent and dependent will be the same.

Note

With interpretation = "series" a subject is declared test positive if both of the tests performed return a positive result. With interpretation = "parallel" a subject is declared test positive if one of the tests performed return a positive result. Intepreting test results in series increases diagnostic specificity. Interpreting test results in parallel increases diagnostic sensitivity.

How do I work out appropriate values for covar? Assume you have two diagnostic tests — an indirect fluorescent antibody test (IFAT) and a polymerase chain reaction (PCR). The diagnostic sensitivity and specificity of the IFAT is 0.784 and 0.951, respectively. The diagnostic sensitivity and specificity of the PCR is 0.926 and 0.979, respectively. These tests are used on a group of individuals known to be disease positive and a group of individuals known to be disease negative. Results for the disease positive group are as follows:
<table>
<thead>
<tr>
<th></th>
<th>IFAT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCR</td>
</tr>
<tr>
<td></td>
<td>Pos</td>
</tr>
<tr>
<td>Pos</td>
<td>134</td>
</tr>
<tr>
<td>Neg</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>138</strong></td>
</tr>
</tbody>
</table>

Results for the disease negative group are as follows:

<table>
<thead>
<tr>
<th></th>
<th>IFAT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCR</td>
</tr>
<tr>
<td></td>
<td>Pos</td>
</tr>
<tr>
<td>Pos</td>
<td>0</td>
</tr>
<tr>
<td>Neg</td>
<td>28</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>28</strong></td>
</tr>
</tbody>
</table>

The observed proportion of disease positive individuals with a positive test result to both tests as \( p_{11} \). For this example \( p_{11} = \frac{134}{176} = 0.761 \).

The observed proportion of disease negative individuals with a negative test result to both tests as \( p_{00} \). For this example \( p_{00} = \frac{534}{574} = 0.930 \).

Covariance for the disease positive group: \( \text{covar}[1] = p_{11} \cdot \text{se}[1] \cdot \text{se}[2] = 0.761 \cdot 0.784 \cdot 0.926 = 0.835 \).

Covariance for the disease negative group: \( \text{covar}[2] = p_{00} \cdot \text{sp}[1] \cdot \text{sp}[2] = 0.930 \cdot 0.951 \cdot 0.979 = -0.001 \).

The covariance for the disease positive group is small. The covariance for the disease negative group is negligible.

References


Examples

```bash
## EXAMPLE 1:
```
## You would like to confirm the absence of disease in a study area. You intend to use two tests: the first has a sensitivity and specificity of 0.90 and 0.80, respectively. The second has a sensitivity and specificity of 0.95 and 0.85, respectively. You need to make sure that an individual that returns a positive test really has disease, so the tests will be interpreted in series (to improve specificity).

## What is the diagnostic sensitivity and specificity of this testing regime?

```r
rsu.dxtest(se = c(0.90,0.95), sp = c(0.80,0.85),
           interpretation = "series", covar = c(0,0))
```

Interpretation of these tests in series returns a diagnostic sensitivity of 0.855 and a diagnostic specificity of 0.97.

## EXAMPLE 2 (from Dohoo, Martin and Stryhn p 113):

An IFAT and PCR are to be used to diagnose infectious salmon anaemia. The diagnostic sensitivity and specificity of the IFAT is 0.784 and 0.951, respectively. The diagnostic sensitivity and specificity of the PCR is 0.926 and 0.979, respectively. It is known that the two tests are dependent, with details of the covariance calculated above. What is the expected sensitivity and specificity if the tests are to be interpreted in parallel?

```r
rsu.dxtest(se = c(0.784,0.926), sp = c(0.951,0.979),
           interpretation = "parallel", covar = c(0.035,-0.001))
```

Interpreting test results in parallel and accounting for the lack of test independence returns a diagnostic sensitivity of 0.949 and diagnostic specificity of 0.930.

---

**rsu.epinf**

**Effective probability of disease**

### Description

Calculates the effective probability of disease (adjusted design prevalence) for each risk group within a population.

### Usage

```r
rsu.epinf(pstar, rr, ppr)
```

### Arguments

- **pstar**
  - scalar, the design prevalence.

- **rr**
  - vector, defining the relative risk values for each strata in the population.

- **ppr**
  - vector of length **rr** defining the population proportions in each strata.
rsu.pfree.equ

Equilibrium probability of disease freedom assuming representative or risk based sampling

Description

Calculates the long-term equilibrium probability of disease freedom and equilibrium prior probability of freedom, after discounting for the probability that disease has been introduced into the population and assuming population sensitivity and probability of introduction are constant over time. It does not specify how long it might take to reach equilibrium.

Usage

rsu.pfree.equ(se.p, p.intro)
Arguments

se.p scalar or vector, the surveillance system (population-level) sensitivity for the given time period.

p.intro scalar or vector of the same length as sep representing the probability of disease introduction for time period.

Value

A list comprised of two elements:

epfree a vector listing the equilibrium probability of disease freedom.

depfree a vector listing the discounted equilibrium probability of disease freedom.

Examples

## EXAMPLE 1:
## The current (ongoing) surveillance system for a given disease in your country has been estimated to have a population sensitivity of 0.60 per time period (one year). Assuming the probability of disease introduction per unit time is 0.02, what is the eventual plateau level for confidence of freedom and how long will it take to reach this level, assuming a prior (starting) confidence of freedom of 0.50?

## Firstly, estimate the equilibrium (plateau) confidence of freedom:

conf.eq <- rsu.pfree.equ(se.p = 0.60, p.intro = 0.02)

conf.eq

## The equilibrium discounted probability of disease freedom is 0.986.

## Next, calculate confidence of freedom over 20 time periods for se.p = 0.60 and p.intro = 0.02:

rval.df <- rsu.pfree.rs (se.p = rep(0.6, times = 20), p.intro = rep(0.02, times = 20), prior = 0.50)

head(rval.df)

## When does the confidence of freedom first reach the equilibrium value (rounded to 3 digits)?

rsep.p <- which(rval.df$pfree >= round(conf.eq$depfree, digits = 3))

rsep.p[1]

## It takes 9 time periods (years) to reach the equilibrium level of 0.986.

## EXAMPLE 2:
## You have been asked to design a surveillance system to detect a given disease in your country. If the probability of disease introduction per unit time is 0.10, what surveillance system sensitivity do you need to be 95% certain that disease is absent based on the testing carried out as
## part of your program?

## Generate a vector of candidate surveillance system sensitivity estimates
## from 0.1 to 0.99:

se.p <- seq(from = 0.10, to = 0.99, by = 0.01)

## Calculate the probability of disease freedom for each of the candidate
## surveillance system sensitivity estimates:

rval.df <- rsu.pfree.equ(se.p = se.p, p.intro = 0.10)
rval.df <- data.frame(se.p = se.p, depfree = rval.df$depfree)
head(rval.df)

## Which of the surveillance system sensitivity estimates returns a
## probability of freedom greater than 0.95?

rsep.p <- rval.df$se.p[rval.df$depfree > 0.95]
rsep.p[1]

## The required surveillance system sensitivity for this program is 0.69.
## Plot the results:

## Not run:
library(ggplot2)

ggplot(data = rval.df, aes(x = se.p, y = depfree)) +
  geom_point() +
  geom_line() +
  scale_x_continuous(limits = c(0, 1),
                    name = "Surveillance system sensitivity") +
  scale_y_continuous(limits = c(0, 1),
                    name = "Equilibrium discounted probability of disease freedom") +
  geom_hline(aes(yintercept = 0.95), linetype = "dashed") +
  geom_vline(aes(xintercept = rsep.p[1]), linetype = "dashed") +
  theme_bw()

## End(Not run)

rsu.pfree.rs calculate probability of freedom for given population sensitivity and probability of introduction

### Description

Calculates the posterior probability (confidence) of disease freedom (negative predictive value) for one or more population sensitivity (se.p) estimates, over one or more time periods.

### Usage

rsu.pfree.rs(se.p, p.intro = 0, prior = 0.5, by.time = TRUE)
Arguments

se.p    scalar, vector or matrix representing the population sensitivity estimates. se.p will be scalar if you’re calculating the posterior probability of disease freedom for a single time period. If se.p is a vector set by.time = TRUE if the se.p estimates are for separate time periods. Set by.time = FALSE if the se.p estimates are variations (iterations) within a single time period. If se.p is a matrix, columns represent consecutive time periods and rows represent multiple se.p estimates per time period.

p.intro scalar, vector or matrix representing the probability of disease introduction per time period. If p.intro is scalar this value is applied across all se.p values and time periods. If p.intro is a vector set by.time = TRUE if the p.intro estimates are for separate time periods. Set by.time = FALSE if the p.intro estimates are variations (iterations) within a single time period. If p.intro is a matrix it should have the same dimensions as se.p with columns representing time periods and rows representing multiple p.intro estimates per time period.

prior    scalar or vector of the same length as the number of rows of se.p representing the prior probability of disease freedom before surveillance.

by.time logical, representing the type of analysis. See details, below.

Details

The by.time argument is used for two specific circumstances.

Use by.time = TRUE if the se.p estimates are a vector of values for consecutive time periods. Use by.time = FALSE if the se.p estimates are a vector of multiple values (iterations) for a single time period.

Use by.times = TRUE if se.p is a symmetrical matrix and p.intro is a vector of values representing the probability of disease introduction over consecutive time periods. Use by.time = FALSE if se.p is a symmetrical matrix (with columns for time periods and rows representing estimates of se.p within each time period) and p.intro is a vector of values corresponding to multiple values for a single time period that are the same across all periods.

Value

A list comprised of six elements:

PFree    The posterior probability of disease freedom.
SeP      The population sensitivity.
PIntro    The probability of disease introduction (as entered by the user).
Discounted prior The discounted prior confidence of disease freedom.
Equilibrium PFree The equilibrium probability of disease freedom.
Equilibrium prior The equilibrium discounted prior probability of disease freedom.
References


Examples

## EXAMPLE 1:
## You have estimated herd-sensitivity for 20 herds for a disease of concern, all returned negative results. What is the confidence of disease freedom for these herds, assuming that based on other data, 20% of herds in the region are estimated to be disease positive?

## Generate 20 herd sensitivity estimates, using random values between 70% and 95%:
herd.sens <- runif(n = 20, min = 0.70, max = 0.95)

## The background herd prevalence is 0.20, so the prior confidence of freedom is 1 - 0.2 = 0.8. For this example we assume the prior is applicable at the time of sampling so p.intro = 0 (the default) and we are carrying out an analysis using multiple estimates of population sensitivities for a single time period so we set by.time = FALSE.

rval.df <- rsu.pfree.rs(se.p = herd.sens, p.intro = 0, prior = 0.80, by.time = FALSE)
rval.df <- data.frame(SeP = rval.df$SeP, PFree = rval.df$PFree)

range(rval.df$SeP)

## The herd-level probability of disease freedom ranges from about 0.93 to 0.99 depending on individual herd level sensitivity values.

## EXAMPLE 2:
## You have analysed 12 months of surveillance data for disease X, to provide 12 monthly estimates of population sensitivity. In addition, based on previous data, the monthly probability of the introduction of disease is estimated to be in the range of 0.005 (0.5%) to 0.02 (2%). The prior confidence of disease freedom is assumed to be 0.5 (i.e. uninformed). What is your level of confidence of disease freedom at the end of the 12 month surveillance period?

## Generate 12, monthly estimates of se.p and p.intro:
pop.sens <- runif(n = 12, min = 0.40, max = 0.70)
pintro <- runif(n = 12, min = 0.005, max = 0.020)

## For this example we're analysing a single population over multiple time
## periods, so we set by.time = TRUE:

```r
rval.df <- rsu.pfree.rs(se.p = pop.sens, p.intro = pintro, prior = 0.50, by.time = TRUE)
rval.df <- data.frame(mnum = 1:12, mchar = seq(as.Date("2020/1/1"), by = "month", length.out = 12), SeP = t(rval.df$SeP), PFree = t(rval.df$PFree))
```

## Plot the probability of disease freedom as a function of time:

```r
plot(x = rval.df$mnum, y = rval.df$PFree, xlim = c(1,12), ylim = c(0,1), xlab = "Month", ylab = "Probability of disease freedom", pch = 16, type = "b", xaxt = "n")
axis(side = 1, at = rval.df$mnum, labels = format(rval.df$mchar, format = "%b"))
abline(h = 0.95, lty = 2)
```

## Not run:

```r
library(ggplot2); library(scales)
ggplot(data = rval.df, aes(x = mchar, y = PFree)) + geom_line(col = "black") + scale_x_date(breaks = date_breaks("1 month"), labels = date_format("%b"), name = "Month") + scale_y_continuous(limits = c(0,1), name = "Probability of disease freedom") + geom_hline(yintercept = 0.95, linetype = "dashed") + theme_bw()
```

## End(Not run)

## The estimated probability of disease freedom (Pfree) increases over time
## from about 0.70 (or less) to >0.99, depending on the actual se.p values
## generated by simulation.

## EXAMPLE 3:
## Extending the above example, instead of a simple deterministic estimate,
## you decide to use simulation to account for uncertainty in the monthly
## se.p and p.intro estimates.

## For simplicity, we generate 1200 random estimates of se.p and coerce them
## into a matrix with 12 columns and 100 rows:

```r
pop.sens <- matrix(runif(n = 1200, min = 0.40, max = 0.70), nrow = 100)
```

## For p.intro we generate a vector of 100 random values, which will then be
## used across all time periods:

```r
pintro <- runif(n = 100, min = 0.005, max = 0.020)
```

## For this example, because se.p is a matrix and p.intro is a vector matching
## one of the dimensions of se.p, by.time is ignored:

```r
rval.df <- rsu.pfree.rs(se.p = pop.sens, p.intro = pintro, prior = 0.5, by.time = TRUE)
rval.df <- data.frame(mnum = 1:12, mchar = seq(as.Date("2020/1/1"), by = "month", length.out = 12), SeP = t(rval.df$SeP), PFree = t(rval.df$PFree))
```

## Plot the probability of disease freedom as a function of time:

```r
plot(x = rval.df$mnum, y = rval.df$PFree, xlim = c(1,12), ylim = c(0,1), xlab = "Month", ylab = "Probability of disease freedom", pch = 16, type = "b", xaxt = "n")
axis(side = 1, at = rval.df$mnum, labels = format(rval.df$mchar, format = "%b"))
abline(h = 0.95, lty = 2)
```

## Not run:

```r
library(ggplot2); library(scales)
ggplot(data = rval.df, aes(x = mchar, y = PFree)) + geom_line(col = "black") + scale_x_date(breaks = date_breaks("1 month"), labels = date_format("%b"), name = "Month") + scale_y_continuous(limits = c(0,1), name = "Probability of disease freedom") + geom_hline(yintercept = 0.95, linetype = "dashed") + theme_bw()
```

## End(Not run)
### rsu.pstar

**Design prevalence back calculation**

**Description**

Calculates design prevalence required for given sample size and desired surveillance system (population-level) sensitivity, assuming representative sampling, imperfect test sensitivity and perfect test specificity.

**Usage**

```r
rsu.pstar(N = NA, n, se.p, se.u)
```
Arguments

N  scalar or vector, integer representing the population size. Use NA if unknown.
n  scalar or vector, integer representing the number of units sampled.
se.p  scalar or vector of the same length as n representing the desired surveillance system (population-level) sensitivity.
se.u  scalar or vector of the same length as n representing the unit sensitivity.

Value

A vector of design prevalence estimates.

References


Examples

## EXAMPLE 1:
## In a study to provide evidence that your country is free of a given disease
## a total of 280 individuals are sampled. Assume a desired surveillance system
## sensitivity of 0.95 and an individual unit diagnostic sensitivity of 0.98.
## If all unit tests return a negative result, what is the maximum prevalence
## if disease is actually present in the population (i.e. what is the design
## prevalence)?
rsu.pstar(N = NA, n = 280, se.p = 0.95, se.u = 0.98)

## If 280 individuals are sampled and tested and each returns a negative test
## result the maximum prevalence (if disease is actually present in the
## population) is 0.011.

## EXAMPLE 2:
## In a study to provide evidence disease freedom a total of 30 individuals
## are sampled from a set of cattle herds. Assume cattle herds in the study
## region range from 100 to 5000 cows. As above, assume a desired surveillance
## system sensitivity of 0.95 and an individuals unit diagnostic sensitivity
## of 0.98. If all 30 unit tests return a negative result, what is the expected
## design prevalence for each herd?
round(rsu.pstar(N = c(100, 500, 1000, 5000), n = 30,
              se.p = 0.95, se.u = 0.98), digits = 3)

## The expected herd level design prevalence ranges from 0.086 (for a 100
## cow herd) to 0.102 (for a 5000 cow herd).
rsu.sep

Probability that the prevalence of disease in a population is less than or equal to a specified design prevalence

Description

Calculates the probability that the prevalence of disease in a population is less than or equal to a specified design prevalence following return of a specified number of negative test results.

Usage

rsu.sep(N, n, pstar, se.u)

Arguments

N  scalar or vector, integer representing the population size.
n  scalar or vector, integer representing the number of units sampled.
pstar  scalar or vector of the same length as n representing the desired design prevalence.
se.u  scalar or vector of the same length as n representing the unit sensitivity.

Value

A vector of the estimated probability that the prevalence of disease in the population is less than or equal to the specified design prevalence.

References


Examples

```r
## EXAMPLE 1:
## The population size in a provincial area is 193,000. In a given two-
## week period 7764 individuals have been tested for COVID-19 using an
## approved PCR test which is believed to have a diagnostic sensitivity of
## 0.85. All individuals have returned a negative result. What is the
## probability that the prevalence of COVID-19 in this population is less
## than or equal to 100 cases per 100,000?
rsu.sep(N = 193000, n = 7764, pstar = 100 / 100000, se.u = 0.85)
```

## If all of the 7764 individuals returned a negative test we can be more than
## 99% confident that the prevalence of COVID-19 in the province is less
rsu.sep.cens

Description
Calculates the surveillance system (population-level) sensitivity for disease detection assuming imperfect test sensitivity, perfect test specificity and when every unit in the population is tested (a census).

Usage
rsu.sep.cens(d = 1, se.u)

Arguments
d scalar integer defining the expected number of infected units in the population (that is, the population size multiplied by the design prevalence).
se.u scalar or vector of numbers between 0 and 1 defining the unit sensitivity of the test.

Value
A vector of surveillance system (population-level) sensitivities.

Examples
## EXAMPLE 1:
## Every animal in a population is to be sampled and tested using a test
## with a diagnostic sensitivity of 0.80. What is the probability that
## disease will be detected if we expect that there are five infected animals
## in the population?
rsu.sep.cens(d = 5, se.u = 0.80)

## The probability that disease will be detected (i.e. the surveillance
rsu.sep.pass

## system sensitivity) is 0.99 (i.e. quite high, even though the sensitivity
## of the test is relatively low).

## EXAMPLE 2:
## Calculate the surveillance system sensitivity assuming every animal in
## populations of size 10, 50, 100, 250 and 500 will be sampled and tested,
## assuming a design prevalence in each population of 0.01 and use of a test
## with a diagnostic sensitivity of 0.92.

rsu.sep.cens(d = ceiling(0.01 * c(10, 50, 100, 250, 500)), se.u = 0.92)

## For the populations comprised of 100 animals or less the surveillance
## system sensitivity is 0.92. For the populations comprised of greater than
## or equal to 250 animals the surveillance system sensitivity is greater
## than 0.99.

---

rsu.sep.pass

### Surveillance system sensitivity assuming passive surveillance and representative sampling within clusters

## Description

Calculates the surveillance system (population-level) sensitivity for detection of disease for a passive surveillance system assuming comprehensive population coverage and sampling of clinical cases within diseased clusters.

## Usage

rsu.sep.pass(N, n, step.p, pstar.c, p.inf.u, se.u)

## Arguments

- **N**
  - scalar or vector of length equal to the number of rows in step.p representing the population size.

- **n**
  - scalar or vector of length equal to the number of rows in step.p representing the number of units tested per cluster.

- **step.p**
  - vector or matrix of detection probabilities (0 to 1) for each step in the detection process. If a vector each value represents a step probability for a single calculation. If a matrix, columns are step probabilities and rows are simulation iterations.

- **pstar.c**
  - scalar (0 to 1) or vector of length equal to the number of rows in step.p representing the cluster-level design prevalence.

- **p.inf.u**
  - scalar (0 to 1) or vector of length equal to the number of rows in step.p representing the probability of disease in sampled and tested units. This is equivalent to the positive predictive value for a given prior probability of infection.

- **se.u**
  - scalar (0 to 1) or vector of length equal to the number of rows in step.p, representing the unit sensitivity.
**Value**

A list comprised of two elements:

- `se.p`: scalar or vector, the estimated surveillance system (population-level) sensitivity of detection.
- `se.c`: scalar or vector, the estimated cluster-level sensitivity of detection.

If `step.p` is a vector, scalars are returned. If `step.p` is a matrix, values are vectors of length equal to the number of rows in `step.p`.

**References**


**Examples**

```r
## EXAMPLE 1:
## A passive surveillance system for disease X operates in your country.
## There are four steps to the diagnostic cascade with detection probabilities
## for each process of 0.10, 0.20, 0.90 and 0.99, respectively. Assuming the
## probability that a unit actually has disease if it is submitted for
## testing is 0.98, the sensitivity of the diagnostic test used at the unit
## level is 0.90, the population is comprised of 1000 clusters (herds),
## five animals from each cluster (herd) are tested and the cluster-level
## design prevalence is 0.01, what is the sensitivity of disease detection
## at the cluster (herd) and population level?

rsu.sep.pass(N = 1000, n = 5, step.p = c(0.10,0.20,0.90,0.99),
          pstar.c = 0.01, p.inf.u = 0.98, se.u = 0.90)

## The sensitivity of disease detection at the cluster (herd) level is 0.018.
## The sensitivity of disease detection at the population level is 0.16.
```

---

**rsu.sep.rb**

*Surveillance system sensitivity assuming risk-based sampling and varying unit sensitivity*

**Description**

Calculates surveillance system (population-level) sensitivity assuming one-stage, risk-based sampling and varying unit sensitivity using either the binomial or hypergeometric methods.

**Usage**

```r
rsu.sep.rb(N, rr, ppr, df, pstar, method = "binomial")
```
Arguments

- **N**: vector of the same length as **rr**, population size estimates for each risk group.
- **rr**: vector of length equal to the number of risk strata, the relative risk values.
- **ppr**: vector of the same length as **rr**, population proportions for each risk group.
- **df**: a dataframe of values for each combination of risk stratum and sensitivity level. Column 1 = risk group index, column 2 = unit sensitivities, column 3 = the sample size for risk group and unit sensitivity.
- **pstar**: scalar, the design prevalence.
- **method**: character string indicating the method to be used. Options are binomial or hypergeometric. See details, below.

Details

If **method** = binomial **N** is ignored and values for **ppr** need to be entered. Conversely, if **method** = hypergeometric, **ppr** is ignored and calculated from **N**.

Value

A list comprised of five elements:

- **sep**: scalar, the population-level sensitivity estimate.
- **epi**: vector, effective probability of infection estimates.
- **adj.risk**: vector, adjusted risks.
- **n**: vector, sample size by risk group
- **se.u**: a vector of the mean sensitivity for each risk group.

Examples

```r
## EXAMPLE 1:
## Calculate the surveillance system sensitivity assuming one-stage risk-based sampling assuming a population comprised of high risk (n = 200 clusters) and low risk (n = 1800 clusters) where the probability of disease in the high risk group is 5 times that of the low risk group.
## Four clusters will be sampled with n = 80, 30, 20 and 30 surveillance units within each cluster tested using a test with diagnostic sensitivity at the surveillance unit level of 0.92, 0.85, 0.92 and 0.85, respectively.
## Assume a design prevalence of 0.01.

gg <- c(1,1,2,2)  
se.u <- c(0.92,0.85,0.92,0.85)  
n <- c(80,30,20,30)  
df <- data.frame(gg, se.u, n)  
rsu.sep.rb(N = c(200,1800), rr = c(5,1), ppr = NA, df = df, pstar = 0.01, method = "hypergeometric")
```
## The expected surveillance system sensitivity is 0.993.

## EXAMPLE 2:
## Recalculate, assuming that we don't know the size of the cluster population
## at risk.

## When the size of the cluster population at risk is unknown we set N = NA
## and enter values for ppr (the proportion of the population in each risk
## group). Assume (from above) that 0.10 of the cluster population are in the
## high risk group and 0.90 are in the low risk group.

rsu.sep.rb(N = NA, rr = c(5,1), ppr = c(0.10,0.90), df = df, pstar = 0.01,
method = "binomial")

## The expected surveillance system sensitivity is 0.980.

---

rsu.sep.rb1rf | Surveillance system sensitivity assuming risk-based sampling on one
risk factor

### Description

Calculates risk-based surveillance system (population-level) sensitivity with a single risk factor,
assuming one-stage risk-based sampling and allowing unit sensitivity to vary among risk strata.

### Usage

```r
rsu.sep.rb1rf(N, n, rr, ppr, pstar, se.u, method = "binomial")
```

### Arguments

- **N**: scalar or vector of the same length as that vector of `rr` defining the population
  size per risk strata. Ignored if `method` = "binomial".
- **n**: scalar or vector of the same length as that vector of `rr` defining the sample size
  per risk strata.
- **rr**: scalar or vector of the same length as that vector of `ppr` defining the relative risk
  values.
- **ppr**: scalar or vector of the same length as that vector of `rr` defining the population
  proportions in each risk strata. Ignored if `method` = "hypergeometric".
- **pstar**: scalar, defining the design prevalence.
- **se.u**: scalar or vector of the same length as that vector of `rr` defining the unit sensitivity
  (which can vary across strata).
- **method**: character string indicating the method to be used. Options are binomial or
  hypergeometric. See details, below.
Details

If `method = binomial` then `N` is ignored and values for `ppr` need to be entered. Conversely, if `method = hypergeometric`, `ppr` is ignored and calculated from `N`.

Value

A list comprised of two elements:

- `se.p` scalar, surveillance system (population-level) sensitivity estimates.
- `epi` vector, effective probability of infection estimates.
- `adj.risk` vector, adjusted relative risk estimates.

Examples

```r
rsu.sep.rb1rf(N = NA, n = c(5,5,5), rr = c(5,3,1), ppr = c(0.10,0.10,0.80), pstar = 0.10, se.u = 0.90, method = "binomial")
```

## The surveillance system sensitivity is 0.94.

```r
rsu.sep.rb1rf(N = c(10,10,80), n = c(5,5,5), rr = c(5,3,1), ppr = NA, pstar = 0.10, se.u = 0.90, method = "hypergeometric")
```

## The surveillance system sensitivity is 0.96, almost identical to that calculated above where the binomial distribution was used to account for not knowing the size of the cluster population at risk.
rsu.sep.rb2rf

Surveillance system sensitivity assuming risk-based sampling on two risk factors

Description

Calculates risk-based surveillance system (population-level) sensitivity with a two risk factors, assuming [one-stage] risk-based sampling and allowing unit sensitivity to vary among risk strata.

Usage

rsu.sep.rb2rf(N, n, rr1, ppr1, rr2, ppr2, pstar, se.u, method = "binomial")

Arguments

N    matrix of population sizes for each risk group. Rows = levels of rr1, columns = levels of rr2.
n    matrix of the number of surveillance units tested in each risk group. Rows = levels of rr1, columns = levels of rr2.
rr1  scalar or vector defining the first set of relative risk values.
ppr1 scalar or vector of the same length as that vector of rr1 defining the population proportions in each of the first risk strata. Proportions must sum to one. Ignored if method = "hypergeometric".
rr2  matrix defining the relative risks for the second risk factor. Rows = levels of rr1, columns = levels of rr2.
ppr2 matrix defining the population proportions in each of the second risk strata. Rows = levels of rr1, columns = levels of rr2. Ignored if method = "hypergeometric".
pstar scalar, defining the design prevalence.
se.u scalar or vector of the same length as that vector of rr1 defining the unit sensitivity (which can vary across strata).
method character string indicating the method to be used. Options are binomial or hypergeometric. See details, below.

Details

If method = binomial N is ignored and values for ppr need to be entered. Conversely, if method = hypergeometric, ppr is ignored and calculated from N.

Value

A list comprised of two elements:

se.p    scalar, surveillance system (population-level) sensitivity estimates.
epi     vector, effective probability of infection estimates.
adj.risk1 vector, adjusted relative risk estimates for the first risk factor.
adj.risk2 vector, adjusted relative risk estimates for the second risk factor.
Examples

## EXAMPLE 1:

A cross-sectional study is to be carried out to confirm the absence of disease using risk based sampling. Assume a design prevalence of 0.01 at the surveillance unit level. Surveillance units are categorised as being either high or low risk with the probability of disease for high risk surveillance units 3 times the probability of disease for low risk units. The proportion of units in each risk group is 0.20 and 0.80, respectively.

Within each of the two risk categories the probability of disease varies with age with younger age groups having four times the risk of disease as older age groups. In the high risk area 10% of the population are young and 90% are old. In the low risk area 30% of the population are young and 70% are old.

The total number of surveillance units in the population is unknown. The numbers of young and old surveillance units tested in the high and low risk groups are 40, 20, 20 and 10, respectively. You intend to use a test with diagnostic sensitivity of 0.80. What is the surveillance system sensitivity?

```
rsu.sep.rb2rf(N = NA, n = rbind(c(40,20), c(20,10)),
              rr1 = c(3,1),
              ppr1 = c(0.20,0.80),
              rr2 = rbind(c(4,1), c(4,1)),
              ppr2 = rbind(c(0.10,0.90), c(0.30,0.70)),
              pstar = 0.01,
              se.u = 0.80, method = "binomial")$se.p
```

The surveillance system sensitivity is 0.93.

## EXAMPLE 2:

This example shows the importance of sampling high risk groups. Take the same scenario as above but switch the relative proportions sampled by risk group --- taking a greater number of samples from the low risk group compared with the high risk group:

```
rsu.sep.rb2rf(N = NA, n = rbind(c(10,20), c(20,40)),
              rr1 = c(3,1),
              ppr1 = c(0.20,0.80),
              rr2 = rbind(c(4,1), c(4,1)),
              ppr2 = rbind(c(0.10,0.90), c(0.30,0.70)),
              pstar = 0.01,
              se.u = 0.80, method = "binomial")$se.p
```

The surveillance system sensitivity is 0.69. Here we've taken exactly the same number of samples as Example 1, but there's a substantial decrease in surveillance system sensitivity because we've concentrated sampling on a low risk group (decreasing our ability to detect disease).
rsu.sep.rb2st

Surveillance system sensitivity assuming risk based, two-stage sampling

Description

Calculates the surveillance system sensitivity for detection of disease assuming risk based, two-stage sampling (sampling of clusters and sampling of units within clusters), imperfect test sensitivity and perfect test specificity. The method allows for a single risk factor at each stage.

Usage

```
rsu.sep.rb2st(H = NA, N = NA, n, rr.c, ppr.c, pstar.c, rr.u, ppr.u, pstar.u, rg, se.u)
```

Arguments

- **H**: scalar, integer representing the total number of clusters in the population. Use NA if unknown.
- **N**: vector, integer representing the number of surveillance units within each cluster. Use NA if unknown.
- **n**: vector, integer representing the number of surveillance units tested within each cluster.
- **rr.c**: cluster level relative risks (vector of length corresponding to the number of risk strata), use rr.c = c(1,1) if a risk factor does not apply.
- **ppr.c**: vector listing the cluster level population proportions for each risk category. Use NA if there are no cluster level risk factors.
- **pstar.c**: scalar, numeric (0 to 1) the cluster-level design prevalence.
- **rr.u**: surveillance unit level relative risks (vector of length corresponding to the number of risk strata), use rr.u = c(1,1) if a risk factor does not apply.
- **ppr.u**: matrix providing the surveillance unit level population proportions for each risk group. One row for each cluster, columns = unit level risk groups, not required if N is provided.
- **pstar.u**: scalar, numeric (0 to 1) the unit-level design prevalence.
- **rg**: vector, listing the risk group (index) for each cluster.
- **se.u**: scalar, numeric (0 to 1), representing the sensitivity of the diagnostic test at the individual surveillance unit level.

Value

A list comprised of:

- **se.p**: the surveillance system (population-level) sensitivity of detection.
- **se.c**: the cluster-level sensitivity of detection.
Examples

## EXAMPLE 1:

You have been asked to provide an assessment of a surveillance program for Actinobacillus hypopneumoniae in pigs. It is known that there are high risk and low risk areas for A. hypopneumoniae in your country with the estimated probability of disease in the high risk area thought to be around 3.5 times that of the probability of disease in the low risk area. It is known that 10% of the 1784 pig herds in the study area are in the high risk area and 90% are in the low risk area.

The risk of A. hypopneumoniae is dependent on age, with adult pigs around five times more likely to be A. hypopneumoniae positive compared with younger (grower) pigs.

Pigs from 20 herds have been sampled: 5 from the low-risk area and 15 from the high-risk area. All of the tested pigs were adults: no grower pigs were tested.

The ELISA for A. hypopneumoniae in pigs has a diagnostic sensitivity of 0.95.

What is the surveillance system sensitivity if we assume a design prevalence of 1 per 100 at the cluster (herd) level and 5 per 100 at the surveillance system unit (pig) level?

# There are 1784 herds in the study area:

H <- 1784

# Twenty of the 1784 herds are sampled. Generate 20 herds of varying size:

set.seed(1234)

hsize <- rlnorm(n = 20, meanlog = log(10), sdlog = log(8))
hsize <- round(hsize + 20, digits = 0)

# Generate a matrix listing the number of growers and finishers in each of the 20 sampled herds. Anywhere between 80% and 95% of the animals in each herd are growers:

set.seed(1234)
pctg <- runif(n = 20, min = 0.80, max = 0.95)
ngrow <- round(pctg * hsize, digits = 0)
nfini <- hsize - ngrow
N <- cbind(ngrow, nfini)

# Generate a matrix listing the number of grower and finisher pigs sampled from each herd:

nsgrow <- rep(0, times = 20)
nsfini <- ifelse(nfini <= 15, nfini, 15)
n <- cbind(nsgrow, nsfini)
# The herd-level design prevalence is 0.01 and the individual pig-level design prevalence is 0.05:

```r
pstar.c <- 0.01
pstar.u <- 0.05
```

# For herds in the high-risk area the probability being A. hyopneumoniae positive is 3.5 times that of herds in the low-risk area. Ninety percent of herds are in the low risk area and 10% are in the high risk area:

```r
rr.c <- c(1,3.5)
ppr.c <- c(0.9,0.1)
```

## We've sampled 5 herds from the low risk area and 15 herds from the high risk area:

```r
rg <- c(rep(1, times = 5), rep(2, times = 15))
```

## For finishers the probability being A. hyopneumoniae positive is 5 times that of growers:

```r
rr.u <- c(1,5)
```

## The diagnostic sensitivity of the A. hyopneumoniae ELISA is 0.95:

```r
se.u <- 0.95
```

```r
rsu.sep.rb2st(H = H, N = N, n = n,
Pstar.c = pstar.c, pstar.u = pstar.u,
rg = rg, rr.c = rr.c, rr.u = rr.u,
ppr.c = ppr.c, ppr.u = NA,
se.u = se.u)
```

## The estimated surveillance system sensitivity of this program is 0.31.

## EXAMPLE 2:
## Repeat these analyses assuming we don't know the total number of pig herds in the population and we have only an estimate of the proportions of growers and finishers in each herd.

## Generate a matrix listing the proportion of growers and finishers in each of the 20 sampled herds:

```r
ppr.u <- cbind(rep(0.9, times = 20), rep(0.1, times = 20))
```

## Set H (the number of clusters) and N (the number of surveillance units within each cluster) to NA:

```r
rsu.sep.rb2st(H = NA, N = NA, n = n,
Pstar.c = pstar.c, pstar.u = pstar.u,
rg = rg, rr.c = rr.c, rr.u = rr.u,
ppr.c = ppr.c, ppr.u = ppr.u,
se.u = se.u)
```
rsu.sep.rbvarse = se.u = se.u)

## The estimated surveillance system sensitivity is 0.20.

rsu.sep.rbvarse  

Surveillance system sensitivity assuming risk based sampling and varying unit sensitivity

Description

Calculates the surveillance system (population-level) sensitivity for detection of disease assuming risk based sampling and varying unit sensitivity.

Usage

rsu.sep.rbvarse(N, rr, df, pstar)

Arguments

N  
scalar integer or vector of integers the same length as rr, representing the population size. Use NA if unknown.

rr  
relative risk values (vector of values corresponding to the number of risk strata).

df  
dataframe of values for each combination of risk stratum and sensitivity level, column 1 = risk group index, column 2 = unit sensitivity, column 3 = n (sample size for risk group and unit sensitivity).

pstar  
scalar representing the design prevalence.

Value

A list comprised of five elements:

sep  
scalar, the population-level sensitivity estimate.

epi  
vector, effective probability of infection estimates.

adj.risk  
vector, adjusted risks.

n  
vector, sample size by risk group

se.u  
a vector of the mean sensitivity for each risk group.

References


Examples

## EXAMPLE 1:

A study has been carried out to detect Johne’s disease in a population of cattle. There are two risk groups (‘high’ and ‘low’) with the risk of disease in the high risk group five times that of the low risk group. The number of animals sampled and unit sensitivity varies by risk group, as detailed below. Assume there number of cattle in the high risk and low risk group is 200 and 1800, respectively.

Calculate the surveillance system sensitivity assuming a design prevalence of 0.01.

```r
rg <- c(1,1,2,2)
se.u <- c(0.92,0.85,0.92,0.85)
n <- c(80,30,20,30)
df <- data.frame(rg = rg, se.u = se.u, n = n)

rsu.sep.rbvarse(N = c(200,1800), rr = c(5,1), df = df, pstar = 0.01)
```

The surveillance system sensitivity is 0.99.

rsu.sep.rs

Surveillance system sensitivity assuming representative sampling

Description

Calculates the surveillance system (population-level) sensitivity for detection of disease assuming representative sampling, imperfect test sensitivity and perfect test specificity using the hypergeometric method if \( N \) is known and the binomial method if \( N \) is unknown.

Usage

```
rsu.sep.rs(N = NA, n, pstar, se.u = 1)
```

Arguments

- **N**: scalar integer or vector of integers the same length as \( n \), representing the population size. Use \( NA \) if unknown.
- **n**: scalar integer or vector of integers representing the number of units tested.
- **pstar**: scalar numeric or vector of numbers the same length as \( n \) representing the design prevalence. See details, below.
- **se.u**: scalar numeric or vector of numbers the same length as \( n \) representing the unit sensitivity.

Value

A vector of surveillance system (population-level) sensitivity estimates.
References


Examples

## EXAMPLE 1:
# Three hundred samples are to be tested from a population of animals to
# confirm the absence of a disease. The total size of the population is
# unknown. Assuming a design prevalence of 0.01 and a test with
# diagnostic sensitivity of 0.95 will be used what is the sensitivity of
# disease detection at the population level?
rsu.sep.rs(N = NA, n = 300, pstar = 0.01, se.u = 0.95)
## The sensitivity of disease detection at the population level is 0.943.

## EXAMPLE 2:
# Thirty animals from five herds ranging in size from 80 to 100 head are to be
# sampled to confirm the absence of a disease. Assuming a design prevalence
# of 0.01 and a test with diagnostic sensitivity of 0.95 will be used, what
# is the sensitivity of disease detection for each herd?
N <- seq(from = 80, to = 100, by = 5)
n <- rep(30, times = length(N))
rsu.sep.rs(N = N, n = n, pstar = 0.01, se.u = 0.95)
## The sensitivity of disease detection for each herd ranges from 0.28 to
## 0.36.

rsu.sep.rs2st

---

rsu.sep.rs2st

**Surveillance system sensitivity assuming representative two-stage sampling**

Description

Calculates the surveillance system sensitivity for detection of disease assuming two-stage sampling (sampling of clusters and sampling of units within clusters), imperfect test sensitivity and perfect test specificity.

Usage

rsu.sep.rs2st(H = NA, N = NA, n, pstar.c, pstar.u, se.u = 1)
Arguments

H scalar, integer representing the total number of clusters in the population. Use NA if unknown.

N vector, integer representing the number of units within each cluster. Use NA if unknown.

n vector, integer representing the number of units tested within each cluster.

pstar.c scalar, numeric (0 to 1) representing the cluster-level design prevalence.

pstar.u scalar, numeric (0 to 1) representing the unit-level design prevalence.

se.u scalar, numeric (0 to 1), representing the sensitivity of the diagnostic test at the individual unit level.

Value

A list comprised of:

se.p the surveillance system (population-level) sensitivity of detection.

se.c the cluster-level sensitivity of detection.

se.u the unit-level sensitivity of detection.

N the number of units within each cluster, as entered by the user.

n the number of units tested within each cluster, as entered by the user.

Note

If pstar.c is not a proportion N must be provided and N must be greater than n.

Examples

```r
## EXAMPLE 1:
## A study is to be conducted to confirm the absence of enzootic bovine
## leukemia disease in your country. Four herds are to be sampled from a
## population of 500 herds. There are 550, 250, 700 and 200 cows in each of
## the four herds. From each of the four herds 30 animals are to be sampled.
## The design prevalence for this study is set to 0.01 at the herd level
## and if a herd is positive for leukemia the individual animal level
## design prevalence is set to 0.10. Assuming a test with diagnostic
## sensitivity of 0.98 will be used, what is the sensitivity of
## disease detection at the population and cluster (herd) level?

rsu.sep.rs2st(H = 500, N = c(550,250,700,200), n = rep(30, times = 4),
       pstar.c = 0.01, pstar.u = 0.10, se.u = 0.98)

## The population level sensitivity of detection is 0.037. The cluster level
## sensitivity of detection ranges from 0.950 to 0.958.
```
rsu.sep.rsfreecalc  
*Surveillance system sensitivity for detection of disease assuming representative sampling and imperfect test sensitivity and specificity.*

**Description**

Calculates the surveillance system (population-level) sensitivity for detection of disease assuming representative sampling and imperfect test sensitivity and specificity.

**Usage**

```r
rsu.sep.rsfreealc(N, n, c = 1, pstar, se.u, sp.u)
```

**Arguments**

- `N` scalar, integer representing the total number of subjects eligible to be sampled. Use NA if unknown.
- `n` scalar, integer representing the total number of subjects sampled.
- `c` scalar, integer representing the cut-point number of positives to classify a cluster as positive. If the number of positives is less than \( c \) the cluster is negative; if the number of positives is greater than or equal to \( c \) the cluster is positive.
- `pstar` scalar, numeric, representing the design prevalence, the hypothetical outcome prevalence to be detected. See details, below.
- `se.u` scalar, numeric (0 to 1) representing the diagnostic sensitivity of the test at the unit level.
- `sp.u` scalar, numeric (0 to 1) representing the diagnostic specificity of the test at the unit level.

**Details**

If a value for \( N \) is entered surveillance system sensitivity is calculated using the hypergeometric distribution. If \( N \) is NA surveillance system sensitivity is calculated using the binomial distribution.

**Value**

A scalar representing the surveillance system (population-level) sensitivity.

**References**


### Examples

#### EXAMPLE 1:
Thirty animals from a herd of 150 are to be tested using a test with diagnostic sensitivity 0.90 and specificity 0.98. What is the surveillance system sensitivity assuming a design prevalence of 0.10 and two or more positive tests will be interpreted as a positive result?

```r
rsu.sep.rsfreecalc(N = 150, n = 30, c = 2, pstar = 0.10, se.u = 0.90, sp.u = 0.98)
```

If a random sample of 30 animals is taken from a population of 150 and a positive test result is defined as two or more individuals returning a positive test, the probability of detecting disease if the population is diseased at a prevalence of 0.10 is 0.87.

#### EXAMPLE 2:
Repeat these calculations assuming herd size is unknown:

```r
rsu.sep.rsfreecalc(N = NA, n = 30, c = 2, pstar = 0.10, se.u = 0.90, sp.u = 0.98)
```

If a random sample of 30 animals is taken from a population of unknown size and a positive test result is defined as two or more individuals returning a positive test, the probability of detecting disease if the population is diseased at a prevalence of 0.10 is 0.85.

---

#### rsu.sep.rsmult

**Surveillance system sensitivity by combining multiple surveillance components**

**Description**

Calculates surveillance system (population-level) sensitivity for multiple components, accounting for lack of independence (overlap) between components.

**Usage**

```r
rsu.sep.rsmult(C = NA, pstar.c, rr, ppr, se.c)
```

**Arguments**

- `C` : scalar integer or vector of the same length as `rr`, representing the population sizes (number of clusters) for each risk group.
- `pstar.c` : scalar (0 to 1) representing the cluster level design prevalence.
- `rr` : vector of length equal to the number of risk strata, representing the cluster relative risks.
ppr vector of the same length as \( r \) representing the cluster level population proportions. Ignored if \( C \) is specified.

se.c surveillance system sensitivity estimates for clusters in each component and corresponding risk group. A list with multiple elements where each element is a dataframe of population sensitivity values from a separate surveillance system component. The first column equals the clusterid, the second column equals the cluster-level risk group index and the third column equals the population sensitivity values.

Value

A list comprised of two elements:

se.p a matrix (or vector if \( C \) is not specified) of population-level (surveillance system) sensitivities (binomial and hypergeometric and adjusted vs unadjusted).

se.component a matrix of adjusted and unadjusted sensitivities for each component.

Examples

```r
## EXAMPLE 1:
## You are working with a population that is comprised of individuals in
## 'high' and 'low' risk area. There are 300 individuals in the high risk
## area and 1200 individuals in the low risk area. The risk of disease for
## those in the high risk area is assumed to be three times that of the low
## risk area.

C <- c(300,1200)
pstar.c <- 0.01
rr <- c(3,1)

## Generate population sensitivity values for clusters in each component of
## the surveillance system. Each of the three dataframes below lists id,
## rg (risk group) and cse (component sensitivity):

comp1 <- data.frame(id = 1:100,
                    rg = c(rep(1, time = 50), rep(2, times = 50)),
                    cse = rep(0.5, times = 100))
comp2 <- data.frame(id = seq(from = 2, to = 120, by = 2),
                    rg = c(rep(1, times = 25), rep(2, times = 35)),
                    cse = runif(n = 60, min = 0.5, max = 0.8))
comp3 <- data.frame(id = seq(from = 5, to = 120, by = 5),
                    rg = c(rep(1, times = 10), rep(2, times = 14)),
                    cse = runif(n = 24, min = 0.7, max = 1))

# Combine the three components into a list:
se.c <- list(comp1, comp2, comp3)

## What is the overall population-level (surveillance system) sensitivity?
```

rsu.sep.rsmult(C = C, pstar.c = pstar.c, rr = rr, ppr = NA, se.c = se.c)
## The overall adjusted system sensitivity (calculated using the binomial
## distribution) is 0.85.

## EXAMPLE 2:
## Assume that you don't know exactly how many individuals are in the high
## and low risk areas but you have a rough estimate that the proportion of
## the population in each area is 0.2 and 0.8, respectively. What is the
## population-level (surveillance system) sensitivity?

ppr <- c(0.20, 0.80)
rsu.sep.rsmult(C = NA, pstar.c = pstar.c, rr = rr, ppr = ppr, se.c = se.c)
## The overall adjusted system sensitivity (calculated using the binomial
## distribution) is 0.85.

---

**rsu.sep.rspool**  
*Surveillance system sensitivity assuming representative sampling, imperfect pooled sensitivity and perfect pooled specificity*

**Description**
Calculates the surveillance system (population-level) sensitivity and specificity for detection of disease assuming representative sampling and allowing for imperfect sensitivity and specificity of the pooled test.

**Usage**

```r
rsu.sep.rspool(r, k, pstar, pse, psp = 1)
```

**Arguments**

- `r`  
  scalar or vector representing the number of pools.

- `k`  
  scalar or vector of the same length as `r` representing the number of individual units that contribute to each pool (i.e. the pool size).

- `pstar`  
  scalar or vector of the same length as `r` representing the design prevalence.

- `pse`  
  scalar or vector of the same length as `r` representing the pool-level sensitivity.

- `psp`  
  scalar or vector of the same length as `r` representing the pool-level specificity.

**Value**
A list comprised of two elements:

- `se.p`  
  scalar or vector, the surveillance system (population-level) sensitivity estimates.

- `sp.p`  
  scalar or vector, the surveillance system (population-level) specificity estimates.
References


Examples

```r
## EXAMPLE 1:
## To confirm your country's disease freedom status you intend to use a test applied at the herd level. The test is expensive so you decide to pool the samples taken from individual herds. If you decide to collect 60 pools, each comprised of samples from five herds what is the sensitivity of disease detection assuming a design prevalence of 0.01 and the sensitivity and specificity of the pooled test equals 1.0?

rsu.sep.rspool(r = 60, k = 5, pstar = 0.01, pse = 1, psp = 1)

## This testing regime returns a population-level sensitivity of disease detection of 0.95.

## EXAMPLE 2:
## Repeat these calculations assuming the sensitivity of the pooled test equals 0.90.

rsu.sep.rspool(r = 60, k = 5, pstar = 0.01, pse = 0.90, psp = 1)

## If the sensitivity of the pooled test equals 0.90 the population-level sensitivity of disease detection is 0.93. How can we improve population-level sensitivity? Answer: include more pools in the study.

rsu.sep.rspool(r = 70, k = 5, pstar = 0.01, pse = 0.90, psp = 1)

## Testing 70 pools, each comprised of samples from 5 herds returns a population-level sensitivity of disease detection of 0.95.
```

---

**rsu.sep.rsvarse**  
*Surveillance system sensitivity assuming representative sampling and varying unit sensitivity*

Description

Calculates the surveillance system (population-level) sensitivity for detection of disease assuming representative sampling and varying unit sensitivity.

Usage

```r
rsu.sep.rsvarse(N = NA, pstar, se.u)
```
Arguments

- **N**: scalar integer or vector of integers the same length as `se.u`, representing the population size. Use NA if unknown.
- **pstar**: scalar representing the design prevalence.
- **se.u**: vector of numbers the same length as `N` representing the individual unit sensitivities.

Value

A vector of surveillance system (population-level) sensitivity estimates.

References


Examples

```r
## EXAMPLE 1:
## A study has been carried out to detect Johne's disease in a population of
cattle. A random sample of 50 herds from a herd population of unknown size
has been selected and, from each selected herd, a variable number of animals
have been tested using faecal culture which is assumed to have a diagnostic
sensitivity in the order of 0.60.

## The number of animals tested in each of the 50 herds is:
set.seed(1234)
ntest <- round(runif(n = 50, min = 10, max = 30), digits = 0)
ntest

## Calculate the herd level sensitivity of disease detection, assuming we've
## been provided with no details of the number of animals in each of the 50
## herds. Assume a within-herd design prevalence of 0.05:
herd.se <- rsu.sep.rs(N = NA, n = ntest, pstar = 0.05, se.u = 0.60)
range(herd.se)

## The herd level sensitivity of detection varies between 0.26 and 0.60.

## Calculate the surveillance system sensitivity assuming a herd-level design
## prevalence of 0.01:
rsu.sep.rsvarse(N = NA, pstar = 0.01, se.u = herd.se)

## The surveillance system sensitivity is 0.20.
```
**rsu.spp.rs**  
*Surveillance system specificity assuming representative sampling*

---

**Description**

Calculates surveillance system (population level) specificity assuming representative sampling and imperfect test specificity.

**Usage**

`rsu.spp.rs(N, n, c = 1, sp.u)`

**Arguments**

- **N**
  - scalar or vector of the same length as that vector of `n` defining the [cluster] population size. Use `NA` if the size of the population not known, or for a more general application see details, below.

- **n**
  - scalar or vector defining the sample size.

- **c**
  - scalar or vector of the same length as that vector of `n` defining the cut-point number of positives to classify a cluster as positive, if the number of positive samples is less than `c` the cluster is declared is negative, if the number of positive samples is greater than `c` the cluster is declared positive.

- **sp.u**
  - scalar (0 to 1) or vector of same length as `n`, the specificity of the diagnostic test at the surveillance unit level.

**Details**

This function calculates population specificity using the hypergeometric distribution if `N` and `c` are provided and the binomial distribution otherwise.

If `N` is provided the number of false positives is fixed, based on `N` and test specificity `sp.u`. This implies that test specificity is a fixed individual-level characteristic (e.g. due to specific cross-reacting infection). If `N` is not supplied, cluster (e.g. herd) specificity is a random binomial function based only on the number of samples and test specificity (i.e. specificity is a function of the test and independent of individual characteristics).

**Value**

A vector of population specificity estimates.

**References**

Examples

## EXAMPLE 1:
## Calculate the surveillance system specificity (i.e. the probability that
## an uninfected population will be correctly identified as negative) if 30
## surveillance units have been tested from a population of 150 using a
## diagnostic test with surveillance unit specificity of 0.90, using a
## cut-point of one or more positives to consider the population positive.

## A specificity of 0.90 means that 9 out of 10 samples from disease-negative
## surveillance units will return a negative result (i.e. one of them will be
## a false positive).

rsu.spp.rs(N = 150, n = 30, c = 1, sp.u = 0.90)

## The surveillance system specificity is 0.03. There is a probability of
## 0.03 that all 30 samples will be negative.

## EXAMPLE 2:
## Now assume we set a cut-point of 6. That is, 6 or more samples have to
## return a positive result for us to declare the population positive:

rsu.spp.rs(N = 150, n = 30, c = 6, sp.u = 0.90)

## The surveillance system specificity is 0.95.

---

### rsu.sspfree.rs

Sample size to achieve a desired probability of disease freedom assuming representative sampling

#### Description

Calculates the required sample size to achieve a desired (posterior) probability of disease freedom assuming representative sampling, imperfect test sensitivity and perfect test specificity.

#### Usage

rsu.sspfree.rs(N = NA, prior, p.intro, pstar, pfree, se.u)

#### Arguments

- **N**
  - scalar integer or vector of integers the same length as n, representing the population size. Use NA if unknown.

- **prior**
  - scalar probability (0 to 1), representing the prior probability that the population is free of disease.

- **p.intro**
  - scalar or vector of the same length as pfree, representing the probability of disease introduction during the next time period.
rsu.sspfree.rs

- `pstar`: scalar numeric or vector of numbers the same length as `pfree` representing the design prevalence.
- `pfree`: scalar numeric or vector of numbers the same length as `pfree` representing the desired probability of disease freedom.
- `se.u`: scalar (0 to 1) representing the sensitivity of the diagnostic test at the surveillance unit level.

**Value**

A list comprised of three elements:

- `n`: a vector listing the required sample sizes.
- `sep`: a vector listing the population sensitivity estimates.
- `adj.prior`: a vector listing the adjusted priors.

**Note**

This function returns the sample size to achieve a desired (posterior) probability of disease freedom. Function `rsu.ssssep.rs` returns the sample size to achieve a desired surveillance system sensitivity.

**References**


**Examples**

```r
## EXAMPLE 1:
## Prior surveillance activities and expert opinion lead you to believe that
## there's a 75% chance that your country is free of disease X. To confirm
## your country's disease freedom status you intend to use a test at the herd
## level which has a diagnostic sensitivity of 0.95. The probability of
## disease introduction during the time period of interest is relatively
## low, say 0.01. How many herds need to be sampled to be 95% confident
## that the country is free of disease X assuming a design prevalence of
## 0.01?

rsu.sspfree.rs(N = NA, prior = 0.75, p.intro = 0.01, pstar = 0.01, 
pfree = 0.95, se.u = 0.95)

## A total of 198 herds need to be sampled to meet the requirements of the
## study.
```
rsu.sssep.rb2st1rf  

Sample size to achieve a desired surveillance system sensitivity assuming risk-based 2-stage sampling on one risk factor at the cluster level

Description

Calculates the sample size to achieve a desired surveillance system sensitivity assuming risk-based 2-stage sampling on one risk factor at the cluster level, imperfect test sensitivity and perfect test specificity.

Usage

rsu.sssep.rb2st1rf(rr, ppr, spr, pstar.c, se.c, pstar.u, se.u, se.p)

Arguments

- **rr**: vector, defining the relative risk values for each strata in the population.
- **ppr**: vector of length `rr` defining the population proportions in each strata.
- **spr**: vector of length `rr` defining the planned number of units to be sampled from each strata.
- **pstar.c**: scalar (either a proportion or integer) defining the cluster level design prevalence.
- **se.c**: scalar proportion, defining the desired cluster level sensitivity.
- **pstar.u**: scalar (either a proportion or integer) defining the surveillance unit level design prevalence.
- **se.u**: scalar (0 to 1) representing the sensitivity of the diagnostic test at the surveillance unit level.
- **se.p**: scalar (0 to 1) representing the desired surveillance system (population-level) sensitivity.

Value

A list comprised of seven elements:

- **n.clusters**: scalar, the total number of clusters to be sampled.
- **n.clusters.per.strata**: a vector of the same length as `rr` listing the numbers of clusters to be sampled from each risk stratum.
- **n.units**: scalar, the total number of units to be sampled.
- **n.units.per.strata**: a vector of the same length of `rr` listing the total numbers of units to be sampled from each risk stratum.
- **n.units.per.cluster**: scalar, the number of units to be sampled from each cluster.
epinf a vector of the same length of rr listing the effective probability of infection for each risk stratum.

adj.risk a vector of the same length of rr listing the adjusted risk values for each risk stratum.

Examples

## EXAMPLE 1:
## A cross-sectional study is to be carried out to confirm the absence of
disease using risk based sampling. The population of interest is comprised
## of individual sampling units managed within clusters.

## Clusters are stratified into 'high', 'medium' and 'low' risk areas
## where the cluster-level risk of disease in the high risk area compared
## with the low risk area is 5 and the cluster-level risk of disease in
## the medium risk area compared with the low risk area is 3.

## The proportions of the population at risk in the high, medium and low
## risk area are 0.10, 0.20 and 0.70, respectively. The proportion of samples
## taken from the high, medium and low risk areas will be 0.40, 0.40 and
## 0.20, respectively.

## You intend to use a test with diagnostic sensitivity of 0.90 and you'd
## like to take a sufficient number of samples to return a cluster-level
## sensitivity of 0.80 and a population-level (system) sensitivity of 0.95.
## How many units need to be sampled to meet the requirements of the study?

rr <- c(5,3,1)
ppr <- c(0.10,0.20,0.70)
spr <- c(0.40,0.40,0.20)

rsu.sssep.rb2st1rf(rr, ppr, spr, pstar.c = 0.01, se.c = 0.80,
pstar.u = 0.10, se.u = 0.90, se.p = 0.95)

## A total of 197 clusters needs to be sampled, 79 from the high risk area,
## 79 from the medium risk area and 39 from the low risk area. A total of
## 18 units should be sampled from each cluster, 3546 units in total.

rsu.sssep.rb2st2rf Sample size to achieve a desired surveillance system sensitivity assuming risk-based 2-stage sampling on two risk factors at either the cluster level, unit level, or both

Description

Calculates the sample size to achieve a desired surveillance system sensitivity assuming risk-based 2-stage sampling on two risk factors at either the cluster level, the unit level or both, imperfect test sensitivity and perfect test specificity.
Usage

rsu.sssep.rb2st2rf(rr.c, ppr.c, spr.c, pstar.c, se.c, rr.u, ppr.u, spr.u, pstar.u, se.u, se.p)

Arguments

rr.c vector, corresponding to the number of risk strata defining the relative risk values at the cluster level.
ppr.c vector of length equal to that of rr.c defining the population proportions at the cluster level.
spr.c vector of length equal to that of rr.c defining the planned surveillance proportions at the cluster level.
pstar.c scalar (either a proportion or integer) defining the cluster level design prevalence.
se.c scalar (proportion), the desired cluster level sensitivity.
rr.u vector, corresponding to the number of risk strata defining the relative risk values at the surveillance unit level.
ppr.u vector of length equal to that of rr.u defining the population proportions at the surveillance unit level.
spr.u vector of length equal to that of rr.u defining the planned surveillance proportions at the surveillance unit level.
pstar.u scalar (either a proportion or integer) defining the surveillance unit level design prevalence.
se.u scalar (0 to 1) representing the sensitivity of the diagnostic test at the surveillance unit level.
se.p scalar (0 to 1) representing the desired surveillance system (population-level) sensitivity.

Value

A list comprised of two elements:

clusters scalar, the total number of clusters to be sampled.
units scalar, the total number of units to sample from each cluster.

Examples

## EXAMPLE 1:
## A cross-sectional study is to be carried out to confirm the absence of
disease using risk based sampling. Assume a design prevalence of 0.02
at the cluster (herd) level and a design prevalence of 0.10 at the
surveillance unit (individual) level. Clusters are categorised as
being either high, medium or low risk with the probability of disease for
clusters in the high and medium risk area 5 and 3 times the probability of
disease in the low risk area. The proportions of clusters in the high,
medium and low risk area are 0.10, 0.20 and 0.70, respectively. The
## proportion of samples from the high, medium and low risk area will be
## 0.40, 0.40 and 0.20, respectively.

## Surveillance units (individuals) are categorised as being either high or
## low risk with the probability of disease for units in the high risk group
## 4 times the probability of disease in the low risk group. The proportions
## of units in the high and low risk groups are 0.10 and 0.90, respectively.
## All of your samples will be taken from units in the high risk group.

## You intend to use a test with diagnostic sensitivity of 0.95 and you’d
## like to take sufficient samples to be 95% certain that you’ve detected
## disease at the population level, 95% certain that you’ve detected disease
## at the cluster level and 95% at the surveillance unit level. How many
## clusters and how many units need to be sampled to meet the requirements
## of the study?

rsu.sssep.rb2st2rf(
  rr.c = c(5,3,1), ppr.c = c(0.1,0.2,0.7), spr.c = c(0.4,0.4,0.2),
  pstar.c = 0.02, se.c = 0.50,
  rr.u = c(4,1), ppr.u = c(0.1, 0.9), spr.u = c(1,0),
  pstar.u = 0.10, se.u = 0.90,
  se.p = 0.95)

## A total of 82 clusters needs to be sampled: 33 from the high risk area,
## 33 from the medium risk area and 16 from the low risk area. A total of
## 9 units should be sampled from each cluster.

### rsu.sssep.rbmrg

Sample size to achieve a desired surveillance system sensitivity assuming risk-based sampling and multiple sensitivity values within risk groups

#### Description

Sample the size to achieve a desired population sensitivity assuming risk-based sampling, multiple sensitivity values within risk groups for each risk group and perfect test specificity.

#### Usage

```r
rsu.sssep.rbmrg(pstar, rr, ppr, spr, spr.rg, se.p, se.u)
```

#### Arguments

- **pstar**: scalar, the design prevalence.
- **rr**: vector of length equal to the number of risk strata, the relative risk values.
- **ppr**: vector of the same length as **rr**, population proportions for each risk group.
- **spr**: vector of the same length as **rr**, the planned surveillance proportions for each risk group.
matrix with rows equal to the number of risk groups and columns equal to the number of sensitivity values (row sums must equal 1), the proportions of samples for each sensitivity value in each risk group.

se.p scalar (0 to 1) representing the desired surveillance system (population-level) sensitivity.

se.u vector (0 to 1) representing the sensitivity of the diagnostic test at the surveillance unit level.

Value

A list comprised of three elements:

n matrix of sample sizes for each risk and sensitivity group.

epi a vector of effective probability of infection estimates.

mean.se a vector of the mean sensitivity for each risk group.

Examples

## EXAMPLE 1:
## You are working with a disease of cattle where the prevalence is believed to vary according to herd type. The risk of disease is 5 times greater in dairy herds and 3 times greater in mixed herds compared with the reference category, beef herds. The distribution of dairy, mixed and beef herds in the population of interest is 0.10, 0.10 and 0.80, respectively.
## You intend to distribute your sampling effort 0.4, 0.4 and 0.2 across dairy, mixed and beef herds, respectively.

## Within each of the three risk groups a single test with a diagnostic sensitivity of 0.95 will be used. How many herds need to be sampled if you want to be 95% certain of detecting disease if it is present in the population at a prevalence of 1% or greater?

## Generate a matrix listing the proportions of samples for each test in each risk group (the number of rows equal the number of risk groups, the number of columns equal the number of tests):

m <- rbind(1,1,1)

rsu.sssep.rbmrg(pstar = 0.01, rr = c(5,3,1), ppr = c(0.1,0.1,0.8),
        spr = c(0.4,0.4,0.2), spr.rg = m, se.p = 0.95, se.u = 0.95)

## A total of 147 herds need to be sampled: 59 dairy, 59 mixed and 29 beef herds.

## EXAMPLE 2:
## Now assume that one of two tests will be used for each herd. The first test has a diagnostic sensitivity of 0.92. The second test has a diagnostic sensitivity of 0.80. The proportion of dairy, mixed and beef herds receiving the first test is 0.80, 0.50 and 0.70, respectively (which means that 0.20, 0.50 and 0.30 receive the second test, respectively).
## Recalculate the sample size.

```r
m <- rbind(c(0.8, 0.2), c(0.5, 0.5), c(0.7, 0.3))
rsu.sssep.rbmrg(pstar = 0.01, rr = c(5, 3, 1), ppr = c(0.1, 0.1, 0.8),
                spr = c(0.4, 0.4, 0.2), spr.rg = m, se.p = 0.95, se.u = c(0.92, 0.80))
```

## A total of 159 herds need to be sampled: 64 dairy, 64 mixed and 31 beef herds.

---

### `rsu.sssep.rbsrg`

**Sample size to achieve a desired surveillance system sensitivity assuming risk-based sampling and a single sensitivity value for each risk group**

Description

Sample the size to achieve a desired population sensitivity assuming risk-based sampling, a single sensitivity value for each risk group and perfect test specificity.

Usage

```r
rsu.sssep.rbsrg(pstar, rr, ppr, spr, se.p, se.u)
```

Arguments

- **pstar**: scalar, representing the design prevalence.
- **rr**: vector, defining the relative risk values for each strata in the population.
- **ppr**: vector of length `rr`, defining the population proportions in each strata.
- **spr**: vector of length `rr` representing the planned surveillance proportion for each strata in the population.
- **se.p**: scalar (0 to 1) representing the desired surveillance system (population-level) sensitivity.
- **se.u**: scalar (0 to 1) or vector of the same length as `rr` representing the sensitivity of the diagnostic test applied at the unit level.

Value

A list of comprised of four elements:

- **n**: a vector listing the required sample sizes for each (risk) strata.
- **total**: scalar, representing the total sample size.
- **epinf**: a vector listing the effective probability of infection estimates.
- **adj.risk**: a vector listing the adjusted risk estimates.
Examples

## EXAMPLE 1:
## A cross-sectional study is to be carried out to confirm the absence of
disease using risk based sampling. Assume a population level design
## prevalence of 0.10 and there are 'high', 'medium' and 'low' risk areas
## where the risk of disease in the high risk area compared with the low risk
## area is 5 and the risk of disease in the medium risk area compared with
## the low risk area is 3. The proportions of the population at risk in the
## high, medium and low risk area are 0.10, 0.10 and 0.80, respectively.
## Half of your samples will be taken from individuals in the high risk area,
## 0.30 from the medium risk area and 0.20 from the low risk area. You intend
## to use a test with diagnostic sensitivity of 0.90 and you'd like to take
## sufficient samples to return a population sensitivity of 0.95. How many
## units need to be sampled to meet the requirements of the study?

rsu.sssep.rbsrg(pstar = 0.10, rr = c(5,3,1), ppr = c(0.10,0.10,0.80),
             spr = c(0.50,0.30,0.20), se.p = 0.95, se.u = 0.90)

## A total of 14 units needs to be sampled to meet the requirements of the
## study: 7 from the high risk area, 5 from the medium risk area and 2 from
## the low risk area.

rsu.sssep.rs Sample size to achieve a desired surveillance system sensitivity assuming representative sampling

Description

Calculates the sample size to achieve a desired surveillance system sensitivity assuming representative sampling for a single risk factor and varying unit sensitivity using the binomial method.

Usage

rsu.sssep.rs(N, pstar, se.p = 0.95, se.u)

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>scalar integer or vector of same length as pstar, representing the population size.</td>
</tr>
<tr>
<td>pstar</td>
<td>a scalar or vector of either proportions (0 to 1) or a positive integers representing the design prevalence. If pstar is an integer represents the number of positive units in the population, and N must be provided.</td>
</tr>
<tr>
<td>se.p</td>
<td>scalar or vector of same length as pstar representing the desired surveillance system (population-level) sensitivity.</td>
</tr>
<tr>
<td>se.u</td>
<td>scalar (0 to 1) or vector of the same length as pstar representing the sensitivity of the diagnostic test at the surveillance unit level.</td>
</tr>
</tbody>
</table>
Value

A vector of required sample sizes.

Note

This function calculates the required sample size using the hypergeometric distribution if N is provided and the binomial distribution otherwise.

This function returns the sample size to achieve a desired surveillance system sensitivity. Function `rsu.sspfree.rs` returns the sample size to achieve a desired (posterior) probability of disease freedom.

References


Examples

```r
## EXAMPLE 1:
## You would like to confirm the absence of disease in a single 1000-cow dairy herd. You expect the prevalence of disease in the herd to be 0.05. You intend to use a single test with a sensitivity of 0.90 and a specificity of 1.00. How many herds need to be sampled if you want to be 95% certain that the prevalence of brucellosis in dairy herds is less than the design prevalence if all tests are negative?

rsu.sssep.rs(N = 1000, pstar = 0.05, se.p = 0.95, se.u = 0.90)

## We need to sample 65 cows.
```

```r
## EXAMPLE 2:
## You would like to confirm the absence of disease in a study area comprised of 5000 herds. If the disease is present you expect the between-herd prevalence to be 0.08. You intend to use two tests: the first has a sensitivity and specificity of 0.90 and 0.80, respectively. The second has a sensitivity and specificity of 0.95 and 0.85, respectively. The two tests will be interpreted in parallel. How many herds should be sampled to be 95% certain that the disease would be detected if it is present in the study area?

## Calculate the sensitivity and specificity of the diagnostic test regime:

test <- rsu.dxttest(se = c(0.90, 0.95), sp = c(0.80, 0.85),
interpretation = "parallel", covar = c(0,0))

## Interpretation of these tests in parallel returns a diagnostic sensitivity of 0.995 and a diagnostic specificity of 0.68.
```
## How many herds should be sampled?

```r
rsu.sssep.rs(N = 5000, pstar = 0.08, se.p = 0.95, se.u = test$se)
```

## If you test 38 herds and all return a negative test you can state that
## you are 95% confident that the disease is absent from the study area.
## The sensitivity of this testing regime is 99%.

## EXAMPLE 3:
## You want to document the absence of Mycoplasma from a 200-sow pig herd.
## Based on your experience and the literature, a minimum of 20% of sows
## would have seroconverted if Mycoplasma were present in the herd. How
## many herds should we sample to be 95% certain that Mycoplasma would
## be detected if it is present if you use a test with perfect sensitivity?

```r
rsu.sssep.rs(N = 200, pstar = 0.20, se.p = 0.95, se.u = 1.00)
```

## If you test 15 sows and all of them test negative you can be 95%
## confident that the prevalence rate of Mycoplasma in the herd is less than
## 20%.

---

**rsu.sssep.rs2st** *Sample size to achieve a desired surveillance system sensitivity assuming two-stage sampling*

### Description

Calculates the required sample size to achieve a desired surveillance system sensitivity assuming two-stage sampling (sampling of clusters and sampling of units within clusters), imperfect test sensitivity and perfect test specificity.

### Usage

```r
rsu.sssep.rs2st(H = NA, N = NA, pstar.c, se.c, pstar.u, se.u, se.p)
```

### Arguments

- **H**: scalar, integer representing the total number of clusters in the population. Use NA if unknown.
- **N**: vector, integer representing the number of units within each cluster. Use NA if unknown.
- **pstar.c**: scalar, numeric (0 to 1) representing the cluster level design prevalence.
- **se.c**: scalar, numeric (0 to 1) representing the required cluster level sensitivity.
- **pstar.u**: scalar, numeric (0 to 1) representing the surveillance unit level design prevalence.
se.u  scalar (0 to 1) representing the sensitivity of the diagnostic test at the surveillance unit level.

se.p  scalar (0 to 1) representing the desired surveillance system (population-level) sensitivity.

Value

A list comprised of two data frames: clusters and units. Data frame clusters lists:

- H  the total number of clusters in the population, as entered by the user.
- nsample  the number of clusters to be sampled.

Data frame units lists:

- N  the number of units within each cluster, as entered by the user.
- nsample  the number of units to be sampled.

References


Examples

```r
## EXAMPLE 1:
## Sampling is to be carried out to support a claim that a country is free
## of bovine brucellosis. We are not certain of the total number of herds
## in the country and we are not certain of the number of cows within each
## herd.

## The design prevalence for this study is set to 0.01 at the herd level and
## if a herd is positive for brucellosis the individual animal level
## design prevalence is set to 0.10. The sensitivity of the diagnostic
## test to be used is 0.95.

## How many herds and how many animals from within each herd
## need to be sampled to be 95% confident of detecting disease at the
## herd and individual animal level?

rsu.sssep.rs2st(H = NA, N = NA, pstar.c = 0.01, se.c = 0.95,
                 pstar.u = 0.10, se.u = 0.95, se.p = 0.95)
```
## A total of 314 herds need to be sampled, 31 cows from each herd.

## EXAMPLE 2:
## Now let's say we know that there are 500 cattle herds in the country and
## we have the results of a recent livestock census providing counts of the
## number of cattle in each herd. How many herds and how many animals from
## within each herd need to be sampled to be 95% confident of detecting
## disease at the herd and individual animal level?

# Generate a vector of herd sizes. The minimum herd size is 25.

set.seed(1234)
hsize <- ceiling(rlnorm(n = 500, meanlog = 1.5, sdlog = 2)) + 25

nsample <- rsu.sssep.rs2st(H = 500, N = hsize, pstar.c = 0.01, se.c = 0.95,
                          pstar.u = 0.10, se.u = 0.95, se.p = 0.95)

nsample$clusters
head(nsample$units)

## A total of 238 of the 500 herds need to be tested. The number of animals
## to sample from the first herd (comprised of 26 animals) is 18.

---

**rsu.sssep.rsfreecalc**  
Sample size to achieve a desired surveillance system sensitivity to detect disease at a specified design prevalence assuming representative sampling, imperfect unit sensitivity and specificity.

### Description
Calculates the sample size to achieve a desired surveillance system sensitivity to detect disease at a specified design prevalence assuming representative sampling, imperfect unit sensitivity and specificity.

### Usage
```
rsu.sssep.rsfreecalc(N, pstar, mse.p, msp.p, se.u, sp.u, method = "hypergeometric", max.ss = 32000)
```

### Arguments
- **N**: scalar, integer representing the total number of subjects eligible to be sampled.
- **pstar**: scalar, numeric, representing the design prevalence, the hypothetical outcome prevalence to be detected. See details, below.
- **mse.p**: scalar, numeric (0 to 1) representing the desired population level sensitivity. See details, below.
msp.p  scalar, numeric (0 to 1) representing the desired population level specificity. See details, below.
se.u  scalar (0 to 1) representing the sensitivity of the diagnostic test at the surveillance unit level.
sp.u  scalar, numeric (0 to 1) representing the specificity of the diagnostic test at the surveillance unit level.
method  a character string indicating the calculation method to use. Options are binomial or hypergeometric.
max.ss  scalar, integer defining the maximum upper limit for required sample size.

Details

Type I error is the probability of rejecting the null hypothesis when in reality it is true. In disease freedom studies this is the situation where you declare a population as disease negative when, in fact, it is actually disease positive. Type I error equals 1 - SeP.

Type II error is the probability of accepting the null hypothesis when in reality it is false. In disease freedom studies this is the situation where you declare a population as disease positive when, in fact, it is actually disease negative. Type II error equals 1 - SpP.

Argument \( pstar \) can be expressed as either a proportion or integer. Where the input value for \( pstar \) is between 0 and 1 the function interprets \( pstar \) as a prevalence. Where the input value for \( pstar \) is an integer greater than 1 the function interprets \( pstar \) as the number of outcome-positive individuals in the population of individuals at risk. A value for design prevalence is then calculated as \( pstar / N \).

Value

A list comprised of two data frames: summary and details. Data frame summary lists:

\( n \) the minimum number of individuals to be sampled.
\( N \) the total number of individuals eligible to be sampled.
\( c \) the cut-point number of positives to achieve the specified surveillance system (population-level) sensitivity and specificity.
\( pstar \) the design prevalence.
\( p1 \) the probability that the population has the outcome of interest at the specified design prevalence.
\( se.p \) the calculated population level sensitivity.
\( sp.p \) the calculated population level specificity.

Data frame details lists:

\( n \) the minimum number of individuals to be sampled.
\( se.p \) the calculated population level sensitivity.
\( sp.p \) the calculated population level specificity.
## References


## Examples

```r
rsu.sssep.rsfreecalc(N = 5000, pstar = 0.05, mse.p = 0.95, msp.p = 0.95, se.u = 0.95, sp.u = 0.98, method = "hypergeometric", max.ss = 32000)
```

A system sensitivity of 95% is achieved with a total sample size of 194 herds, assuming a cut-point of 7 or more positive herds are required to return a positive survey result.

---

### rsu.sssep.rspool

**Sample size to achieve a desired surveillance system sensitivity using pooled samples assuming representative sampling**

**Description**

Calculates the required sample size to achieve a desired surveillance system sensitivity assuming representative sampling, imperfect pooled test sensitivity and imperfect pooled test specificity.

**Usage**

```r
rsu.sssep.rspool(k, pstar, pse, psp, se.p)
```

**Arguments**

- `k`: scalar or vector of the same length as `sep` representing the number of individual units that contribute to each pool (i.e. the pool size).
- `pstar`: scalar or vector of the same length as `sep` representing the design prevalence.
- `pse`: scalar or vector of the same length as `sep` representing the pool-level sensitivity.
rsu.sssep.rspool

\texttt{psp} \\
scalar or vector of the same length as \texttt{sep} representing the pool-level specificity.

\texttt{se.p} \\
scalar or vector (0 to 1) representing the desired surveillance system (population-level) sensitivity.

\textbf{Value}

A vector of required sample sizes.

\textbf{References}


\textbf{Examples}

\texttt{## EXAMPLE 1:}
\texttt{## To confirm your country's disease freedom status you intend to use a test}
\texttt{## applied at the herd level. The test is expensive so you decide to pool the}
\texttt{## samples taken from individual herds. How many pooled samples of size 5 are}
\texttt{## required to be 95\% confident that you will have detected disease if}
\texttt{## 1\% of herds are disease-positive? Assume a diagnostic sensitivity and}
\texttt{## specificity of 0.90 and 0.95 for the pooled testing regime.}

\texttt{rsu.sssep.rspool(k = 5, pstar = 0.01, pse = 0.90, psp = 0.95, se.p = 0.95)}

\texttt{## A total of 32 pools (each comprised a samples from 5 herds) need to be}
\texttt{## tested.}
## Index

<table>
<thead>
<tr>
<th><em>datasets</em></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>epi.epidural, 52</td>
<td></td>
</tr>
<tr>
<td>epi.incin, 54</td>
<td></td>
</tr>
<tr>
<td>epi.SClip, 86</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><em>htest</em></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>epi.occ, 74</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><em>methods</em></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>rsu.adjrisk, 146</td>
<td></td>
</tr>
<tr>
<td>rsu.dxtest, 149</td>
<td></td>
</tr>
<tr>
<td>rsu.epinf, 151</td>
<td></td>
</tr>
<tr>
<td>rsu.pfree.equ, 152</td>
<td></td>
</tr>
<tr>
<td>rsu.pfree.rs, 154</td>
<td></td>
</tr>
<tr>
<td>rsu.pstar, 158</td>
<td></td>
</tr>
<tr>
<td>rsu.sep, 160</td>
<td></td>
</tr>
<tr>
<td>rsu.sep.cens, 161</td>
<td></td>
</tr>
<tr>
<td>rsu.sep.pass, 162</td>
<td></td>
</tr>
<tr>
<td>rsu.sep.rb, 163</td>
<td></td>
</tr>
<tr>
<td>rsu.sep.rb1rf, 165</td>
<td></td>
</tr>
<tr>
<td>rsu.sep.rb2rf, 167</td>
<td></td>
</tr>
<tr>
<td>rsu.sep.rb2st, 169</td>
<td></td>
</tr>
<tr>
<td>rsu.sep.rbvarse, 172</td>
<td></td>
</tr>
<tr>
<td>rsu.sep.rs, 173</td>
<td></td>
</tr>
<tr>
<td>rsu.sep.rs2st, 174</td>
<td></td>
</tr>
<tr>
<td>rsu.sep.rcsult, 177</td>
<td></td>
</tr>
<tr>
<td>rsu.sep.respool, 179</td>
<td></td>
</tr>
<tr>
<td>rsu.sep.rsvarse, 180</td>
<td></td>
</tr>
<tr>
<td>rsu.spp.rs, 182</td>
<td></td>
</tr>
<tr>
<td>rsu.ssfree.rs, 183</td>
<td></td>
</tr>
<tr>
<td>rsu.ssssep.rb2st1rf, 185</td>
<td></td>
</tr>
<tr>
<td>rsu.ssssep.rb2st2rf, 186</td>
<td></td>
</tr>
<tr>
<td>rsu.ssssep.rbmg, 188</td>
<td></td>
</tr>
<tr>
<td>rsu.ssssep.rbmg, 190</td>
<td></td>
</tr>
<tr>
<td>rsu.ssssep.rs, 191</td>
<td></td>
</tr>
<tr>
<td>rsu.ssssep.respool, 197</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><em>univar</em></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>epi.2by2, 4</td>
<td></td>
</tr>
<tr>
<td>epi.about, 16</td>
<td></td>
</tr>
<tr>
<td>epi.asc, 20</td>
<td></td>
</tr>
<tr>
<td>epi.betabuster, 21</td>
<td></td>
</tr>
</tbody>
</table>
rsu.sep.cens, 20, 161
rsu.sep.pass, 20, 162
rsu.sep.rb, 19, 163
rsu.sep.rb1rf, 19, 165
rsu.sep.rb2rf, 19, 167
rsu.sep.rb2st, 19, 169
rsu.sep.rbvarse, 19, 172
rsu.sep.rs, 19, 173
rsu.sep.rs2st, 19, 174
rsu.sep.rsfreecalc, 19, 176
rsu.sep.rsmult, 19, 177
rsu.sep.rspool, 19, 179
rsu.sep.rsvarse, 19, 180
rsu.spp.rs, 19, 182
rsu.sspfree.rs, 19, 183, 192
rsu.sssep.rb2st1rf, 19, 185
rsu.sssep.rb2st2rf, 19, 186
rsu.sssep.rbmrg, 19, 188
rsu.sssep.rbsrg, 19, 190
rsu.sssep.rs, 19, 184, 191
rsu.sssep.rs2st, 19, 193
rsu.sssep.rsfreecalc, 19, 195
rsu.sssep.rspool, 19, 197

summary.epi.2by2 (epi.2by2), 4
summary.epi.occc (epi.occc), 74
summary.epi.tests (epi.tests), 142