Package ‘epiR’

February 25, 2016

Version 0.9-74
Date 2016-01-24
Title Tools for the Analysis of Epidemiological Data
Author Mark Stevenson <mark.stevenson1@unimelb.edu.au> with contributions from Telmo Nunes, Cord Heuer, Jonathon Marshall, Javier Sanchez, Ron Thornton, Jeno Reiczigel, Jim Robison-Cox, Paola Sebastiani, Peter Solymos, Kazuki Yoshida, Geoff Jones, Sarah Pirikahu, Simon Firestone and Ryan Kyle
Maintainer Mark Stevenson <mark.stevenson1@unimelb.edu.au>
Description Tools for the analysis of epidemiological 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, and computing confidence intervals around incidence risk and incidence rate estimates. Miscellaneous functions for use in meta-analysis, diagnostic test interpretation, and sample size calculations.
Depends R (>= 3.0.0), survival
Imports BiasedUrn, methods
Suggests MASS (>= 3.1-20)
License GPL (>= 2)
URL http://fvas.unimelb.edu.au/veam
NeedsCompilation no
Repository CRAN
Date/Publication 2016-02-25 23:38:55

R topics documented:

epi.2by2 ......................................................... 3
epi.about ......................................................... 12
epi.asc .......................................................... 13
epi.betabuster .................................................... 14
epi.bohning ...................................................... 16
epi.ccc .......................................................... 17
epi.cluster1size ............................................... 19
2

R topics documented:

epi.cluster2size ....................................................... 20
epi.clustersize .......................................................... 25
epi.conf ................................................................. 27
epi.convgrid ............................................................. 32
epi.cp ....................................................................... 33
epi.cpresids ............................................................ 34
epi.descriptives ......................................................... 35
epi.detectsize ........................................................... 36
epi.dgamma ............................................................... 38
epi.directadj ............................................................ 39
epi.dms ................................................................. 41
epi.dsl ................................................................. 42
epi.edr ................................................................. 44
epi.empbayes ........................................................... 45
epi.epidural .............................................................. 47
epi.equivb .............................................................. 47
epi.equivc .............................................................. 50
epi.herdtest ............................................................. 52
epi.incin ............................................................... 53
epi.indirectadj .......................................................... 54
epi.insthaz ............................................................. 57
epi.interaction ......................................................... 58
epi.iv ................................................................. 60
epi.kappa ............................................................... 62
epi.ltd ................................................................. 66
epi.mh ................................................................. 67
epi.nomogram ......................................................... 69
epi.noninfb ............................................................. 70
epi.noninfc ............................................................. 72
epi.occc ............................................................... 75
epi.offset ............................................................. 76
epi.pooled ............................................................. 77
epi.popsize ............................................................ 78
epi.prcc ............................................................. 79
epi.prev ............................................................. 80
epi.RtoBUGS ........................................................... 82
epi.SClip ............................................................. 83
epi.simplesize ......................................................... 84
epi.smd ............................................................. 86
epi.stratasize .......................................................... 88
epi.studysize ......................................................... 90
epi.tests ............................................................ 95

Index .................................................. 99
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. Multiple strata may be represented by additional rows of count data. 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,
   homogeneity = "breslow.day", outcome = "as.columns")

## S3 method for class 'epi.2by2'
print(x, ...)

## S3 method for class 'epi.2by2'
summary(object, ...)

Arguments

dat an object of class table containing the individual cell frequencies. See the examples, below, for details.

method a character string indicating the experimental design on which the tabular data has been based. Options are cohort.count, cohort.time, case.control, or cross-sectional.

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

units multiplier for prevalence and incidence estimates.

homogeneity a character string indicating the type of homogeneity test to perform. Options are breslow.day or woolf.

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:

   ______  ______  ______  ______
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 specifying the experimental design on which the tabular data has been based.
- `n.strata` number of strata.
- `conf.level` magnitude of the returned confidence intervals.
- `massoc` a list comprised of the computed measures of association. See below for details.
- `tab` a data frame comprised of of the contingency table data.

When method equals `cohort.time` and `outcome = as.columns` the following measures of association and effect are returned:

- **RR** Wald and score confidence intervals for the incidence risk ratios for each strata. Wald 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 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.
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.

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

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

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

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

Mantel-Haenszel chi-squared test.

Test of homogeneity of the individual strata incidence risk ratios.

Test of homogeneity of the individual strata odds ratios.

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

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.

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.

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

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

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

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

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

Mantel-Haenszel chi-squared test.

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

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.

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.
<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PARisk</strong></td>
<td>Wald and Pirikahu confidence intervals for the population attributable risk for each strata. Wald and Pirikahu confidence intervals for the crude population attributable risk.</td>
</tr>
<tr>
<td><strong>AFest</strong></td>
<td>Wald confidence intervals for the estimated attributable fraction for each strata. Wald confidence intervals for the crude estimated attributable fraction.</td>
</tr>
<tr>
<td><strong>PAFest</strong></td>
<td>Wald confidence intervals for the population estimated attributable fraction for each strata. Wald confidence intervals for the crude population estimated attributable fraction.</td>
</tr>
<tr>
<td><strong>chisq.strata</strong></td>
<td>chi-squared test for difference in exposed and non-exposed proportions for each strata.</td>
</tr>
<tr>
<td><strong>chisq.crude</strong></td>
<td>chi-squared test for difference in exposed and non-exposed proportions across all strata.</td>
</tr>
<tr>
<td><strong>chisq.mh</strong></td>
<td>Mantel-Haenszel chi-squared test.</td>
</tr>
<tr>
<td><strong>OR.homog</strong></td>
<td>test of homogeneity of the individual strata odds ratios.</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PR</strong></td>
<td>Wald and score confidence intervals for the prevalence ratios for each strata. Wald confidence intervals for the crude prevalence ratio. Wald confidence interval for the Mantel-Haenszel adjusted prevalence ratio.</td>
</tr>
<tr>
<td><strong>OR</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>ARisk</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>PARisk</strong></td>
<td>Wald and Pirikahu confidence intervals for the population attributable risk for each strata. Wald and Pirikahu confidence intervals for the crude population attributable risk.</td>
</tr>
<tr>
<td><strong>AFRisk</strong></td>
<td>Wald confidence intervals for the attributable fraction for each strata. Wald confidence intervals for the crude attributable fraction.</td>
</tr>
<tr>
<td><strong>PAFRisk</strong></td>
<td>Wald confidence intervals for the population attributable fraction for each strata. Wald confidence intervals for the crude population attributable fraction.</td>
</tr>
<tr>
<td><strong>chisq.strata</strong></td>
<td>chi-squared test for difference in exposed and non-exposed proportions for each strata.</td>
</tr>
<tr>
<td><strong>chisq.crude</strong></td>
<td>chi-squared test for difference in exposed and non-exposed proportions across all strata.</td>
</tr>
<tr>
<td><strong>chisq.mh</strong></td>
<td>Mantel-Haenszel chi-squared test.</td>
</tr>
<tr>
<td><strong>PR.homog</strong></td>
<td>test of homogeneity of the individual strata prevalence ratios.</td>
</tr>
<tr>
<td><strong>OR.homog</strong></td>
<td>test of homogeneity of the individual strata odds ratios.</td>
</tr>
</tbody>
</table>
Note

Measures of strength 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 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 disease in the exposed group attributable to exposure.

Measures of total effect 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 disease in the population that may be attributed to exposure. The population attributable fraction is the proportion of the disease 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 odds ratio are calculated using Wald (Wald 1943), score (Miettinen and Nurminen 1985) and maximum likelihood methods (Fleiss et al. 2003). 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 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.

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 tests of homogeneity of the odds (risk) ratio where homogeneity = "breslow.day" and homogeneity = "woolf" 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) and
Kazuki Yoshida (Brigham and Women’s Hospital, Boston Massachusetts, USA). Thanks to Ian Dohoo for numerous helpful suggestions to improve the documentation for this function.

References


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

```r
## 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 <- matrix(c(13,2163,5,3349), nrow = 2, byrow = TRUE)
rownames(dat) <- c("DF+", "DF-")); colnames(dat) <- c("FUS+", "FUS-")); dat
epi.2by2(dat = as.table(dat), method = "cross-sectional",
  conf.level = 0.95, units = 100, homogeneity = "breslow.day",
  outcome = "as.columns")

## Outcome variable (FUS) as rows:
dat <- matrix(c(13,5,2163,3349), nrow = 2, byrow = TRUE)
rownames(dat) <- c("FUS+", "FUS-")); colnames(dat) <- c("DF+", "DF-")); dat
epi.2by2(dat = as.table(dat), method = "cross-sectional",
  conf.level = 0.95, units = 100, homogeneity = "breslow.day",
  outcome = "as.rows")

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

## Attributable fraction:
## In DCF exposed cats, 75% of FUS is attributable to DCF (95% CI 30% to
## 91%).

## Population attributable fraction:
## Fifty-four percent of FUS cases in the cat population are attributable
## to DCF (95% CI 4% to 78%).
```
## Example 2:
This example shows how the table function can be used to pass data to epi.2by2. Here we use the birthwgt data from the MASS package.

```r
library(MASS)
dat1 <- birthwt; head(dat1)

# Generate a table of cell frequencies. First set the levels of the outcome and the exposure so the frequencies in the 2 by 2 table come out in the conventional format:
dat1$low <- factor(dat1$low, levels = c(1,0))
dat1$smoke <- factor(dat1$smoke, levels = c(1,0))
dat1$race <- factor(dat1$race, levels = c(1,2,3))

# Generate the 2 by 2 table. Exposure (rows) = smoke. Outcome (columns) = low.
tab1 <- table(dat1$smoke, dat1$low, dnn = c("Smoke", "Low BW"))
print(tab1)

# Compute the incidence risk ratio and other measures of association:
epi.2by2(dat = tab1, method = "cohort.count",
         conf.level = 0.95, units = 100, homogeneity = "breslow.day",
         outcome = "as.columns")

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

# Now stratify by race:
tab2 <- table(dat1$smoke, dat1$low, dat1$race,
              dnn = c("Smoke", "Low BW", "Race"))
print(tab2)

# Compute the crude odds ratio, the Mantel-Haenszel adjusted odds ratio and other measures of association:
rval <- epi.2by2(dat = tab2, method = "cohort.count",
                 conf.level = 0.95, units = 100, homogeneity = "breslow.day",
                 outcome = "as.columns")
print(rval)

# After adjusting for the confounding effect of race, the odds of having a low birth weight child for smokers is 2.15 (95% CI 1.29 to 3.58) 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:
rval$massoc$ARisk.mh.green
rval$massoc$ARisk.mh.wald

# Now turn tab2 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:
data2 <- data.frame(tab2)
print(data2)

## Re-format data2 (a summary count data frame) into tabular format using the
## xtabs function:
tab3 <- xtabs(Freq ~ Smoke + Low.BW + Race, data = data2)
print(tab3)

# tab3 can now be passed to epi.2by2:
rv <- epi.2by2(data = tab3, method = "cohort.count",
              conf.level = 0.95, units = 100, homogeneity = "breslow.day",
              outcome = "as.columns")
print(rv)

## 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-Haensel 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?
rv$massOC.OR.crude.cfield
rv$massOC.OR.crude.mle
rv$massOC.OR.crude.score

## Cornfield: 2.20 (95% CI 1.07 to 3.79)
## Maximum likelihood: 2.01 (1.03 to 3.96)
## Score: 2.20 (95% CI 2.84 to 5.17)

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

## Not run:
library(ggplot2); library(scales)
nstrata <- 1:dim(tab3)[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.l <- c(rv$massOC.OR.mh$lower, rv$massOC.OR.strata.cfield$lower)
or.u <- c(rv$massOC.OR.mh$upper, rv$massOC.OR.strata.cfield$upper)
or.p <- c(rv$massOC.OR.mh$est, rv$massOC.OR.strata.cfield$est)
dat <- data.frame(y.at, y.lab, or.p, or.l, or.u)
p <- ggplot(dat, aes(or.p, y.at))
p + 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) +
### Example

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 <- as.table(matrix(c(136,22050,1709,127650), nrow = 2, byrow = TRUE))
rval <- epi.2by2(dat = dat, method = "cohort.time", conf.level = 0.90,
    units = 1000, homogeneity = "breslow.day", outcome = "as.columns")
summary(rval)$ARate.strata.wald
```

The incidence rate of cancer was 7.22 cases per 1000 person-years less in the blind, compared with those who were not blind but had severe visual impairment (90% CI 6.20 to 8.24 cases per 1000 person-years).

```r
round(summary(rval)$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, 90% CI 0.40 to 0.53).

---

### epi.about

**The library epiR: summary information**

#### Description

Tools for the analysis of epidemiological data.

#### Usage

```r
epi.about()
```

#### Details

The most recent version of the epiR package can be obtained from: [http://fvas.unimelb.edu.au/veam](http://fvas.unimelb.edu.au/veam)

#### Author(s)

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

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

Telmo Nunes, UISEE/DE TSA, 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:

\[
\begin{pmatrix}
3 & 4
\end{pmatrix}
\]
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.

---

### 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

```r
# 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 <- epi.betabuster(mode = 0.90, conf = 0.80, greaterthan = TRUE,
                        x = 0.75, conf.level = 0.95, max.shape1 = 100, step = 0.001)
rval$shape1; rval$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$shape1 - 1; r

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

# Density plot of the estimated beta distribution:
plot(seq(from = 0, to = 1, by = 0.001),
```

...
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

obs the observed number of cases in each area.
exp the expected number of cases in each area.
alpha alpha level to be used for the test of significance. Must be a single number between 0 and 1.

Value

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

References


Examples

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

**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.

**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`.
- `nmissing`: a count of the number of measurement pairs ignored due to missingness.
References


See Also

epi.ccc

Examples

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

## Introduce some missing values:
method1[50] <- NA
method2[75] <- NA

tmp.ccc <- epi.ccc(method1, method2, ci = "z-transform",
conf.level = 0.95)

lab <- paste("CCC: ", round(tmp.ccc$rho.c[,1], digits = 2), " (95% CI ",
round(tmp.ccc$rho.c[,2], digits = 2), ", ",
round(tmp.ccc$rho.c[,3], digits = 2), ")", sep = "")
z <- lm(method2 ~ method1)

par(pty = "s")
plot(method1, method2, xlab = c(0, 5), ylab = c(0,5), xlim = "Method 1",
ylab = "Method 2", pch = 16)
abline(a = 0, b = 1, lty = 2)
abline(z, lty = 1)
```
epi.cluster1size

Sample size under under one-stage cluster sampling

Description

Returns the required number of clusters to be sampled using a one-stage cluster sampling strategy.

Usage

epi.cluster1size(n, mean, var, epsilon.r, method = "mean", conf.level = 0.95)

Arguments

n integer, representing the total number of clusters in the population.
mean number, representing the population mean of the variable of interest.
var number, representing the population variance of the variable of interest.
epsilon.r the maximum relative difference between our estimate and the unknown population value.
method a character string indicating the method to be used. Options are total, mean or mean.per.unit.
conf.level scalar, defining the level of confidence in the computed result.
Value

Returns an integer defining the required number of clusters to be sampled.

References


Examples

## A survey to estimate the total number of residents over 65 years of age that require the services of a nurse is to be carried out. There are five housing complexes in the study area and we expect that there might be a total of around 34 residents meeting this criteria (variance 6.8). We would like the estimated sample size to provide us with an estimate that is within 10% of the true value. How many housing complexes (clusters) should be sampled?

epi.cluster2size(n = 5, mean = 34, var = 6.8, epsilon.r = 0.10, method = "total", conf.level = 0.999)

## We would need to sample 3 housing complexes to meet the specifications for this study.

### epi.cluster2size

**Sample size under two-stage cluster sampling**

Description

Returns the required number of clusters to be sampled using a two-stage cluster sampling strategy.

Usage

epi.cluster2size(nbar, R, n, mean, sigma2.x, sigma2.y, sigma2.xy, epsilon.r, method = "mean", conf.level = 0.95)

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>nbar</td>
<td>integer, representing the total number of listing units to be selected from each cluster.</td>
</tr>
<tr>
<td>R</td>
<td>scalar, representing an estimate of the unknown population prevalence to be estimated. Only used when method = &quot;proportion&quot;.</td>
</tr>
<tr>
<td>n</td>
<td>vector of length two, specifying the total number of clusters in the population and the total number of listing units within each cluster, respectively.</td>
</tr>
<tr>
<td>mean</td>
<td>vector of length two, specifying the mean of the variable of interest at the cluster level and listing unit level, respectively.</td>
</tr>
</tbody>
</table>
sigma2.x  vector of length two, specifying the variance of the [denominator] variable of interest at the cluster level and listing unit level, respectively.
sigma2.y  vector of length two, specifying the variance of the numerator variable of interest at the cluster level and listing unit level, respectively. See details. Only used when method = "proportion".
sigma2.xy  vector of length two, specifying the covariance at the cluster level and listing unit level, respectively. Only used when method = "proportion".
epsilon.r  the maximum relative difference between the estimate and the unknown population value.
method    a character string indicating the method to be used. Options are total, mean or proportion.
conf.level scalar, defining the level of confidence in the computed result.

Details

In simple two-stage cluster sampling the number of listing units to be selected from each cluster is determined on the basis of cost and on the basis of the relative sizes of the first- and second-stage variance components. Once the number of listing units is fixed we might then wish to determine the total number of clusters to be sampled to be confident of obtaining estimates that reflect the true population value.

Value

Returns an integer defining the required number of clusters to be sampled.

References


Examples

```r
## 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). How many health centres should be sampled?

tn <- c(5, 3); tmean <- c(84, 28); tsigma2.x <- c(567, 160)
epi.cluster2size(nbar = 2, n = tn, mean = tmean, sigma2.x = tsigma2.x,
                   sigma2.y = NA, sigma2.xy = NA, epsilon.r = 0.3, method = "mean",
                   conf.level = 0.95)
```

## Three health centres need to be sampled to meet the survey
## Specifications.

## EXAMPLE 2 (from Levy and Lemeshow p 294):
## Same scenario as above, but this time we want to estimate the proportion
## of patients referred to a general practitioner from each clinic. As before,
## we want to be 95% confident that our estimate of the proportion of referred
## patients is within 36% of the true population value. We expect that
## approximately 36% of patients are referred.

## On page 295 Levy and Lemeshow state that the parameters \( \sigma_2.x \), \( \sigma_2.y \)
## and \( \sigma_2.xy \) are rarely known in advance and must be either estimated
## or guessed from experience or intuition. In this example (for
## demonstration) we use the actual patient data to calculate \( \sigma_2.x \),
## \( \sigma_2.y \) and \( \sigma_2.xy \).

## Nurse-level data. The following code reproduces Table 10.4 of Levy and
## Lemeshow (page 293).
 clinic <- rep(1:5, each = 3)
 nurse <- 1:15
 Xij <- c(58,44,18,42,53,10,13,18,37,16,32,10,25,23,23)
 Yij <- c(5,6,6,3,19,2,12,6,30,5,14,4,17,9,14)
 ssudat <- data.frame(clinic, nurse, Xij, Yij)

Xbar <- by(data = ssudat$xij, INDICES = ssudat$clinic, FUN = mean)
ssudat$xbar <- rep(Xbar, each = 3)
Ybar <- by(data = ssudat$yij, INDICES = ssudat$clinic, FUN = mean)
ssudat$ybar <- rep(Ybar, each = 3)

ssudat$xij.Xbar <- (ssudat$xij - ssudat$xbar)^2
ssudat$yij.Ybar <- (ssudat$yij - ssudat$ybar)^2
ssudat$XY <- (ssudat$xij - ssudat$xbar) * (ssudat$yij - ssudat$ybar)

## Collapse the nurse-level data (created above) to the clinic level.
## The following code reproduces Table 10.3 of Levy and Lemeshow (page 292).
 clinic <- as.vector(by(ssudat$clinic, INDICES = ssudat$clinic, FUN = min))
 Xi <- as.vector(by(ssudat$xij, INDICES = ssudat$clinic, FUN = sum))
 Yi <- as.vector(by(ssudat$yij, INDICES = ssudat$clinic, FUN = sum))
 psudat <- data.frame(clinic, Xi, Yi)

psudat$Xi.Xbar <- (psudat$Xi - mean(psudat$Xi))^2
psudat$Yi.Ybar <- (psudat$Yi - mean(psudat$Yi))^2
psudat$XY <- (psudat$Xi - mean(psudat$Xi)) * (psudat$Yi - mean(psudat$Yi))

## Number of primary and secondary sampling units:
 npsu <- nrow(psudat)
 nssu <- mean(by(ssudat$nurse, INDICES = ssudat$clinic, FUN = length))
 tn <- c(npsu, nssu)

## Mean of X at primary sampling unit and secondary sampling unit level:
 tmean <- c(mean(psudat$xi), mean(ssudat$xij))

## Variance of number of patients seen:
epi.cluster2size

```r
tsigma2.x <- c(mean(psudat$Xi_bar), mean(ssudat$Xij.Xbar))

## Variance of number of patients referred:

tsigma2.y <- c(mean(psudat$Yi_bar), mean(ssudat$Yij.Ybar))
tsigma2.xy <- c(mean(psudat$XY), mean(ssudat$XY))

epi.cluster2size(nbar = 2, R = 0.36, n = tn, mean = tmean, 
sigma2.x = tsigma2.x, sigma2.y = tsigma2.y, sigma2.xy = tsigma2.xy, 
epsilon.r = 0.3, method = "proportion", conf.level = 0.95)

## Two health centres need to be sampled to meet the survey 
## specifications.

## EXAMPLE 3: 
## We want to determine the prevalence of brucellosis in dairy cattle in a 
## country comprised of 20 provinces. The number of dairy herds per province 
## ranges from 50 to 1200. Herd size ranges from 25 to 900. We suspect that 
## the prevalence of brucellosis-positive herds across the entire country 
## is around 10%. We suspect that there are a small number of provinces 
## with a relatively high individual cow-level prevalence of disease 
## (thought to be between 40% and 80%). How many herds should be sampled 
## from each province if we want our estimate of prevalence to be within 
## 30% of the true population value?

epi.simplesize(N = 1200, Vsq = NA, Py = 0.10, epsilon.r = 0.30, 
method = "proportion", conf.level = 0.95)

## A total of 234 herds should be sampled from each province.

## Next we work out the number of provinces that need to be sampled. 
## Again, we would like to be 95% confident that our estimate is within 
## 30% of the true population value. Simulate some data to derive appropriate 
## estimates of sigma2.x, sigma2.y and sigma2.xy.

## Number of herds per province:
npsu <- 20

nherds.p <- as.integer(runif(n = npsu, min = 50, max = 1200))

## Mean herd size per province:

hsize.p <- as.integer(runif(n = npsu, min = 25, max = 900))

## Simulate estimates of the cow-level prevalence of brucellosis in each 
## province. Here we generate an equal mix of 'low' and 'high' brucellosis 
## prevalence provinces:

prev.p <- c(runif(n = 15, min = 0, max = 0.05), 
runif(n = 5, min = 0.40, max = 0.80))

## Generate some data:

prov <- c(); herd <- c();
Xij <- c(); Yij <- c();
Xbar <- c(); Ybar <- c();
Xij.Xbar <- c(); Yij.Ybar <- c();
for(i in 1:npsu){
    ## Province identifiers:
    tprov <- rep(i, times = nherds.p[i])
    prov <- c(prov, tprov)

    ## Herd identifiers:
    therd <- 1:nherds.p[i]
    herd <- c(herd, therd)

    ## Number of cows in each of the herds in this province:
    txij <- as.integer(rnorm(n = nherds.p[i], meanlog = log(hsize.p[i]),
                            sdlog = 0.5))
    txbar <- mean(txij)
    txij.Xbar <- (txij - txbar)^2
    Xij <- c(Xij, txij)
    Xbar <- c(Xbar, rep(txbar, times = nherds.p[i]))
    Xij.Xbar <- c(Xij.Xbar, txij.Xbar)

    ## Number of brucellosis-positive cows in each herd:
    tyij <- c()
    for(j in 1:nherds.p[i]){  
        ttyij <- rbinom(n = 1, size = txij[j], prob = prev.p[i])
        tyij <- c(tyij, ttyij)
    }  
    tybar <- mean(tyij)
    tyij.Ybar <- (tyij - tybar)^2
    Yij <- c(Yij, tyij)
    Ybar <- c(Ybar, rep(tybar, times = nherds.p[i]))
    Yij.Ybar <- c(Yij.Ybar, tyij.Ybar)
}

ssudat <- data.frame(prov, herd, Xij, Yij, Xbar, Ybar, Xij.Xbar, Yij.Ybar)
ssudat$XY <- (ssudat$Xij - ssudat$Xbar) * (ssudat$Yij - ssudat$Ybar)

## Collapse the herd-level data (created above) to the province level:
prov <- as.vector(by(ssudat$prov, INDICES = ssudat$prov, FUN = min))
Xij <- as.vector(by(ssudat$Xij, INDICES = ssudat$prov, FUN = sum))
Yij <- as.vector(by(ssudat$Yij, INDICES = ssudat$prov, FUN = sum))
psudat <- data.frame(prov, Xij, Yij)

psudat$Xbar <- (psudat$Xij - mean(psudat$Xij))^2
psudat$Ybar <- (psudat$Yij - mean(psudat$Yij))^2
psudat$XY <- (psudat$Xij - mean(psudat$Xij)) * (psudat$Yij - mean(psudat$Yij))

## Number of primary and secondary sampling units:

npsu <- nrow(psudat)
nssu <- round(mean(by(ssudat$herd, INDICES = ssudat$prov, FUN = length)),
               digits = 0)
tn <- c(npsu, nssu)

## Mean of X at primary sampling unit and secondary sampling unit level:
tmean <- c(mean(psudat$Xij), mean(ssudat$Xij))
## Variance of herd size:
\[ t \sigma_{\text{RN}}^2 = c(\text{mean}(\text{psudat}$X_{i}$), \text{mean}(\text{ssudat}$X_{ij}$)) \]

## Variance of number of brucellosis-positive cows:
\[ t \sigma_{\text{RN}}^2 = c(\text{mean}(\text{psudat}$Y_{i}$), \text{mean}(\text{ssudat}$Y_{ij}$)) \]
\[ t \sigma_{\text{RN}x} = c(\text{mean}(\text{psudat}$XY$), \text{mean}(\text{ssudat}$XY$)) \]

## Finally, calculate the number of provinces to be sampled:
\[ t R = \frac{\text{sum}(\text{psudat}$Y_{i}$)}{\text{sum}(\text{psudat}$X_{i}$)} \]

epi.cluster2size(nbar = 234, R = tR, n = tn, mean = tmean, 
\[ \text{sigma2}$x \text{=} t \sigma_{\text{RN}}^2, \text{sigma2}$y \text{=} t \sigma_{\text{RN}}^2, \text{sigma2}$xy \text{=} t \sigma_{\text{RN}x}, \epsilon_{\text{mean}}.r = 0.3, \text{method} \text{=} \text{"proportion"}, \text{conf.level} = 0.95) \]

## Four provinces (sampling 234 herds from each) are required to be 95% confident that our estimate of the individual animal prevalence of brucellosis is within 30% of the true population value.

### epi.clustersize

**Sample size for cluster-sample surveys**

**Description**

Estimates the number of clusters to be sampled using a cluster-sample design.

**Usage**

```r
epi.clustersize(p, b, rho, epsilon.r, conf.level = 0.95)
```

**Arguments**

- `p` the estimated prevalence of the outcome in the population.
- `b` the number of units sampled per cluster.
- `rho` the intra-cluster correlation, a measure of the variation between clusters compared with the variation within clusters.
- `epsilon.r` scalar, the acceptable relative error.
- `conf.level` scalar, defining the level of confidence in the computed result.

**Value**

A list containing the following:

- `clusters` the estimated number of clusters to be sampled.
- `units` the total number of units to be sampled.
- `design` the design effect.
Note

The intra-cluster correlation (\( \rho \)) will be higher for those situations where the between-cluster variation is greater than within-cluster variation. The design effect depends on \( \rho \) and \( b \) (the number of units sampled per cluster). Note that \( b \) is the number of units sampled per cluster, not the total number of units per cluster. \( \rho = \frac{(D - 1)}{(b - 1)} \).

Design effects of 2, 4, and 7 can be used to estimate \( \rho \) when intra-cluster correlation is low, medium, and high (respectively). A design effect of 7.5 should be used when the intra-cluster correlation is unknown.

References


Examples

```r
## EXAMPLE 1:
## The expected prevalence of disease in a population of cattle is 0.10.
## We wish to conduct a survey, sampling 50 animals per farm. No data
## are available to provide an estimate of \( \rho \), though we suspect
## the intra-cluster correlation for this disease to be moderate.
## We wish to be 95% certain of being within 10% of the true population
## prevalence of disease. How many herds should be sampled?

p <- 0.10; b <- 50; D <- 4
rho <- (D - 1) / (b - 1)
epi.clustersize(p = 0.10, b = 50, rho = rho, epsilon.r = 0.10,
               conf.level = 0.95)

## We need to sample 278 herds (13900 samples in total).

## EXAMPLE 2 (from Bennett et al. 1991):
## A cross-sectional study is to be carried out to determine the prevalence
## of a given disease in a population using a two-stage cluster design. We
## estimate prevalence to be 0.20 and we expect \( \rho \) to be in the order of 0.02.
## We want to take sufficient samples to be 95% certain that our estimate of
## prevalence is within 5% of the true population value (that is, a relative
## error of 0.05 / 0.20 = 0.25). Assuming 20 responses from each cluster,
## how many clusters do we need to be sample?

epi.clustersize(p = 0.20, b = 20, rho = 0.02, epsilon.r = 0.25,
               conf.level = 0.95)

## We need to sample 18 clusters (360 samples in total).
```
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 and smr.
- **method**: a character string indicating the method to use. Where ctype = "inc.risk" or ctype = "prevalence" options are exact, wilson and fleiss. 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 require a two-column matrix; the first column specifies the number of positives, the second column specifies the number of negatives. Method smr 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, prevalence or inc.rate 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:
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

## EXAMPLE 1:
dat <- rnorm(n = 100, mean = 0, sd = 1)
epi.conf(dat, 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 <- data.frame(group = group, val = val)
epi.conf(dat, 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.
before <- c(148, 142, 136, 138, 140, 132, 144, 128, 170, 162, 150, 138, 154, 126, 116)
after <- c(152, 152, 134, 148, 144, 136, 144, 150, 146, 174, 162, 162, 146, 156, 132, 126)
dat <- data.frame(before, after)
dat <- data.frame(cbind(before, after))
epi.conf(dat, 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 <- as.matrix(cbind(pos, neg))
round(epi.conf(dat, 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 onychomycosis.
## Out of 56 patients receiving terbinafine, 5 patients experienced
## adverse effects. Out of 29 patients receiving a placebo, none experienced
## adverse effects.

grpl <- matrix(cbind(5, 51), ncol = 2)
## 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 thallium-201 stress test. The 41 sets of images were
## classified as normal or not by the core thallium 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>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ischaemic</td>
<td>Normal</td>
<td>Total</td>
</tr>
<tr>
<td>--------</td>
<td>------------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>Ischaemic</td>
<td>14</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>Normal</td>
<td>0</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>27</td>
<td>41</td>
</tr>
</tbody>
</table>

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

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

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

The design effect is 12.6. Now calculate the proportion protected:

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

\[
epi.conf(dat, ctype = "prevalence", method = "fleiss", N = 1000000,
\]

\[
\text{design = D, conf.level = 0.95)}
\]

The estimated proportion of the population protected is 0.57 (95% CI
0.53 -- 0.60). If we had mistakenly assumed that data were a simple random
sample the confidence interval 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 easting and northing coordinates.

**Usage**

`epi.convgrid(os.refs)`

**Arguments**

- `os.refs` 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.refs` the method returns a NA.
Examples

```r
os.refs <- c("SJ505585","SJ488573","SJ652636")
epi.convgrid(os.refs)
```

---

### Description

Extract the set of unique patterns from a set of covariates.

### 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 covariates.

### Details

A covariate pattern is a unique combination of values of predictor variables. For example, if a model contains two dichotomous predictors, there will be four covariate patterns possible: (1, 1), (1, 0), (0, 1), and (0, 0). This function extracts the *n* unique covariate patterns from a data set comprised of *i* observations, labelling them from 1 to *n*. A vector of length *m* is also returned, listing the covariate pattern identifier for each observation.

### Value

A list containing the following:

- `cov.pattern`:
  
a data frame with columns: `id` the unique covariate patterns, `n` the number of occasions each of the listed covariate pattern appears in the data, and the unique covariate combinations.

- `id`:
  
a vector listing the covariate pattern identifier for each observation.

### References

Examples

```r
## 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 <- data.frame(obs, v1, v2)

dat.glm <- glm(obs ~ v1 + v2, family = binomial, data = dat)
dat.mf <- model.frame(dat.glm)

## Covariate pattern:
epi.cp(dat.mf[-1])

## 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

```r
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`: an `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

Descriptive statistics

Computes descriptive statistics from a vector of numbers.

Usage

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

Arguments

dat vector for which descriptive statistics will be calculated.
conf.level magnitude of the returned confidence intervals. Must be a single number between 0 and 1.

Value

A list containing the following:

arithmetic n number of observations, mean arithmetic mean, sd arithmetic standard deviation, q25 25th 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, 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.
**Examples**

```r
id <- 1:1000
tmp <- rnorm(1000, mean = 0, sd = 1)
id <- sample(id, size = 20)
tmp[id] <- NA
epi.descriptives(tmp, conf.level = 0.95)
```

**Description**

Estimates the required sample size to detect disease. The method adjusts sample size estimates on the basis of test sensitivity and specificity and can account for series and parallel test interpretation.

**Usage**

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

**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.

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

- **finite.correction**
  logical, should a finite correction factor be applied?
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

The finite correction factor reduces the variance of the sample as the sample size approaches the population size. As a rule of thumb, set `finite.correction = TRUE` when the sample size is greater than 5% of the population size.

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 $p_{111} = se1 \times se2$. The covariance between test results for the disease-negative group is $p_{000} = sp1 \times sp2$.

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 0.80. How many samples should we take to be 95% certain
## that, if all tests are negative, the disease is not present?

epi.detectsize(N = 1000, prev = 0.05, se = 0.90, sp = 0.80, interpretation =
"series", covar = c(0,0), conf.level = 0.95, finite.correction = TRUE)

## We need to sample 59 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.detectsize(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),
conf.level = 0.95, finite.correction = TRUE)
```
## We need to sample 31 cows from 38 herds (a total of 1178 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.detectsize(N = 200, prev = 0.20, se = 1.00, sp = 1.00, conf.level = 0.95,
finite.correction = TRUE)
```

## If you test 12 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.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
## 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 guess at the precision of the heterogeneity term would be:

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

## This can be translated into a 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 may be 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(z, fz, type = "l", ylab = "Probability density of tau")
```

Description

Compute directly adjusted incidence rates.

Usage

```r
epi.directadj(obs, pop, 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 covariates to be adjusted for (e.g. age class, gender). The sum of each row will equal the total number of events for each stratum. If there are no covariates to be adjusted for `obs` will be a one column matrix. The rows the `obs` matrix must be named with the appropriate strata names. The columns of `obs` must be named with the appropriate level identifiers for the covariate. See the example, below.
pop  a matrix representing population time at risk. Rows represent strata (e.g. region); columns represent the covariates to be adjusted for (e.g. age class, gender). The sum of each row will equal the total population time at risk within each stratum. If there are no covariates pop will be a one column matrix. The rows the pop matrix must be named with the appropriate strata names. The columns of pop must be named with the appropriate level identifiers for the covariate. 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 the covariate(s).

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

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 returned by this function (based on Fay and Feuer, 1997) are appropriate for incidence risk data. In this situation the argument pop represents the size of the population at risk (instead of population time at risk).

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

```r
## 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. For the matrix titled pop
## the numbers represent dog-years at risk. The following data were
## obtained for male and female dogs:

obs <- matrix(data = c(15,46,53,16), nrow = 2, byrow = TRUE, 
dimnames = list(c("ED","GL"), c("M","F")))

pop <- matrix(data = c(48,212,180,71), nrow = 2, byrow = TRUE, 
dimnames = list(c("ED","GL"), c("M","F")))

## Compute directly adjusted seroprevalence estimates, using a standard
## population with equal numbers of male and female dogs:

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

epi.directadj(obs, pop, std, units = 1, conf.level = 0.95)

#> $crude
#> strata  cov est lower upper
#> 1 ED    M 0.312500 0.1749039 0.5154212
#> 2 GL    M 0.294444 0.2205591 0.3851406
#> 3 ED    F 0.216981 0.1588575 0.2894224
#> 4 GL    F 0.225352 0.1288082 0.3659577

#> $crude.strata
#> strata  est lower upper
#> 1 ED    0.2346154 0.1794622 0.3013733
#> 2 GL    0.2749004 0.2138889 0.3479040

#> $adj.strata
#> strata  est lower upper
#> 1 ED    0.2647406 0.1860047 0.3692766
#> 2 GL    0.2598983 0.1964162 0.3406224

## The confounding effect of sex has been removed by the gender-adjusted
## incidence rate estimates.
```

---

**epi.dms**

*Decimal degrees and degrees, minutes and seconds conversion*

**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 <- matrix(c(41, 38, 7.836, -40, 40, 27.921),
    byrow = TRUE, nrow = 2)
epi.dms(dat)

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

epi.dsl Mixed-effects meta-analysis of binary outcomes using the DerSimon-
nian 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, the standard error of 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, the standard error of 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, the standard error of the DerSimonian and Laird summary odds ratio, 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, the standard error of the DerSimonian and Laird summary risk ratio, 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

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 \hspace{1cm} \textit{Estimated dissemination ratio}

Description

Computes estimated dissemination ratio on the basis of a vector of numbers (usually counts of incident cases identified on each day of an epidemic).

Usage

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

Arguments

dat \hspace{1cm} \text{a numeric vector listing the number of incident cases for each day of an epidemic.}

n \hspace{1cm} \text{scalar, defining the number of days to be used when computing the estimated dissemination ratio.}

conf.level \hspace{1cm} \text{magnitude of the returned confidence interval. Must be a single number between 0 and 1.}

nsim \hspace{1cm} \text{scalar, defining the number of simulations to be used for the confidence interval calculations.}

na.zero \hspace{1cm} \text{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 \( n_{\text{sim}} \) 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

```
set.seed(123)
dat <- rpois(n = 50, lambda = 2)
edr.04 <- epi.edr(dat, n = 4, conf.level = 0.95, nsim = 99, na.zero = TRUE)

## Plot:
plot(1:50, 1:50, xlim = c(0,25), ylim = c(0,10), xlab = "Days",
ylab = "Estimated dissemination ratio", type = "n", main = "")
lines(1:50, edr.04[,1], type = "l", lwd = 2, lty = 1, col = "blue")
lines(1:50, edr.04[,2], type = "l", lwd = 1, lty = 2, col = "blue")
lines(1:50, edr.04[,3], type = "l", lwd = 1, lty = 2, col = "blue")
```

Description

Computes empirical Bayes estimates of observed event counts using the method of moments.
Usage

epi.empbayes(obs, pop)

Arguments

obs   a vector representing the observed event counts in each unit of interest.

pop   a vector representing the population count in each unit of interest.

Details

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 \((obs + \nu)/(pop + \alpha)\).

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

data(epi.SClip)
obs <- epi.SClip$cases; pop <- epi.SClip$population

est <- epi.empbayes(obs, pop)
empbayes.prop <- (obs + est[4]) / (pop + est[3])
raw.prop <- (obs) / (pop)
rank <- rank(raw.prop)
dat <- data.frame(rank, raw.prop, empbayes.prop)

plot(dat$rank, dat$raw.prop, type = "n", xlab = "Rank", ylab = "Risk")
points(dat$rank, dat$raw.prop, pch = 16, col = "red")
points(dat$rank, dat$empbayes.prop, pch = 16, col = "blue")
legend(x = "topleft", legend = c("Raw estimate", "Bayes adjusted estimate"),
       col = c("red","blue"), pch = c(16,16), bty = "n")
**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**


---

**Description**

Estimate the sample size for a parallel equivalence trial, binary outcomes.

**Usage**

```r
epi.equivb(treat, control, delta, n, r = 1, power, 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 a proportion.
- **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.
- **alpha**: scalar, defining the desired alpha level.

Value

A list containing one or more of the following:

- **n_treat**: the required number of study subject in the treatment group.
- **n_control**: the required number of study subject in the control group.
- **n_total**: the total number of study subjects required.

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: Ps and Pn.

With a superiority trial we specify the maximum acceptable difference between Pn and Ps 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 Pn and Ps 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 Pn and Ps 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>
</table>


Superiority trial: H1 is that the new treatment is better than the standard treatment.
Equivalence trial: H1 is that the new treatment is not too different from the standard treatment.
Non-inferiority trial: H1 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.equivb(treat = 0.50, control = 0.50, delta = 0.20, n = NA, r = 1,
           power = 0.80, 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

Computes the sample size for a parallel equivalence trial with a continuous outcome variable.

Usage

epi.equivc(treat, control, sd, delta, n, r = 1, power, alpha)

Arguments

treat the expected mean of the outcome of interest in the treatment group.
treatment the expected mean of the outcome of interest in the control group.
std the expected population standard deviation of the outcome of interest.
delta the equivalence limit, expressed as a proportion.
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.
alpha scalar, defining the desired alpha level.

Value

A list containing one or more of the following:
n.treat the required number of study subject in the treatment group.
n.control the required number of study subject in the control group.
n.total the total number of study subjects required.

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: Ps and Pn.

With a superiority trial we specify the maximum acceptable difference between Pn and Ps as delta. The null hypothesis is H0: Pn - Ps <= delta and the alternative hypothesis is H1: Pn - Ps > 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 Pn and Ps as delta. The null hypothesis is H0: |Ps - Pn| >= delta and the alternative hypothesis is H1: |Ps - Pn| < 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 \texttt{epi.equivb}.

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):
## It is anticipated that patients on a particular drug have a mean diastolic
## blood pressure of 96 mmHg, as against 94 mmHg on an alternative. It is also
## anticipated that the standard deviation of diastolic BP is approximately
## 8 mmHg. If one wishes to confirm that the difference is likely to be less
## than 5 mmHg, that is, one wishes to show equivalence, how many patients
## are need to be enrolled in the trial? We assume 80% power and
## 95% significance.

epi.equivc(treat = 94, control = 96, sd = 8, delta = 5, n = NA,
  r = 1, power = 0.80, 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% (i.e., \( \delta = 0.05 \)) change of LDL is considered of clinically meaningful difference. Assume the standard 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 \( 0.21 \) units.

\[
eqi.equivc(\text{treat} = 0.20, \text{control} = 0.21, \text{sd} = 0.10, \text{delta} = 0.05, n = \text{NA}, r = 1, \text{power} = 0.80, \text{alpha} = 0.05)
\]

A total of 216 subjects need to be enrolled in the trial, 108 in the treatment group and 108 in the control group.

## Example 2 (cont.):

Suppose only 150 subjects were enrolled in the trial, 75 in the treatment group and 75 in the control group. What is the estimated study power?

\[
eqi.equivc(\text{treat} = 0.20, \text{control} = 0.21, \text{sd} = 0.10, \text{delta} = 0.05, n = 150, r = 1, \text{power} = \text{NA}, \text{alpha} = 0.05)
\]

With only 150 subjects the estimated study power is 0.58.

---

**epi.herdtest**

*Estimate herd test characteristics*

### 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

\[
\text{epi.herdtest}(\text{se}, \text{sp}, \text{p}, \text{N}, \text{n}, \text{k})
\]

### 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: \( AP_{\text{pos}} \) the probability of obtaining a positive test, \( AP_{\text{neg}} \) the probability of obtaining a negative test, \( HSE \) the estimated group (herd) sensitivity, and \( HSP \) 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
```r
## EXAMPLE 1:  
## We wish 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.herdtset(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 covariate 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 covariates 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
Examples

## 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):

obs <- matrix(data = c(58, 130), nrow = 2, byrow = TRUE,
               dimnames = list(c("A", "B"), ""))
pop <- matrix(data = c(1000, 2000), nrow = 2, byrow = TRUE,
               dimnames = list(c("A", "B"), ""))
std <- 0.060
epi.indirectadj(obs = obs, pop = pop, std = std, 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:

obs <- matrix(data = c(58, 130), nrow = 2, byrow = TRUE,
               dimnames = list(c("A", "B"), ""))
pop <- matrix(data = c(550, 450, 500, 1500), nrow = 2, byrow = TRUE,
              dimnames = list(c("A", "B"), c("Beef", "Dairy")))
std <- matrix(data = c(0.025, 0.085, 0.060), nrow = 1, byrow = TRUE,
              dimnames = list("", c("Beef", "Dairy", "Total")))
epi.indirectadj(obs = obs, pop = pop, std = std, units = 100,
                 conf.level = 0.95)

## > $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 52 1.1153846 0.8461538 1.403846
## > B 130 140 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

**Instantaneous hazard computed on the basis of a Kaplan-Meier survival function**

**Description**

Compute the 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 failure, equivalent to the proportion of the population failing per unit time.

**Value**

A data frame with three elements: `time` the observed failure times, `est` the proportion of the population failing per unit time, `lower` the lower bounds of the confidence interval, and `upper` the upper bounds of the confidence interval.

**References**


Examples

```r
require(survival)
ovarian.km <- survfit(Surv(futime,fustat) ~ 1, data = ovarian)

ovarian.haz <- epi.insthaz(ovarian.km, conf.level = 0.95)
plot(ovarian.haz$time, ovarian.haz$est, xlab = "Days",
ylab = "Instantaneous hazard", type = "b", pch = 16)
```

Description

Computes the relative excess risk due to interaction, the proportion of disease among those with both exposures attributable to interaction, and the synergy index for case-control data. Confidence interval calculations are based on those described by Hosmer and Lemeshow (1992).

Usage

```r
epi.interaction(model, coeff, type = c("RERI", "APAB", "S"), conf.level = 0.95)
```

Arguments

- `model`: an object of class `glm`, `coxph` or `mle2`.
- `coeff`: a vector specifying the position of the two coefficients of their interaction term in the model.
- `type`: character string defining the type of analysis to be run. Options are `RERI` the relative excess risk due to interaction, `APAB` 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 is defined as a departure from additivity of effects in epidemiologic studies. This function calculates three indices defined by Rothman (1998): (1) the relative excess risk due to interaction (RERI), (2) the proportion of disease among those with both exposures that is attributable to their interaction (AP[AB]), and (3) the synergy index (S). The synergy index measures the interaction between two risk factors expressed as the ratio of the relative excess risk for the combined effect of the risk factors and the sum of the relative excess risks for each separate effect of the two risk factors. In the absence of interaction both RERI and AP[AB] = 0 and S = 1.

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 further details.

RERI, APAB and S can be used to assess additive interaction when the odds ratio estimates the risk ratio. However, it is recognised that odds ratios from case-control studies are not designed to directly estimate the risk or rate ratio (and only do so well when the outcome of interest is rare).

Value

A data frame listing:

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

References


Examples

```r
## Data from Rothman and Keller (1972) evaluating the effect of joint exposure
## to alcohol and tobacco 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 <- data.frame(alc, smk, can)

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

## 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
```
## 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

```r
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, the standard error of 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, the standard error of 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, the standard error of the inverse variance summary odds ratio, the lower and upper bounds of the confidence interval of the inverse variance summary odds ratio.
- **RR.summary**: the inverse variance summary risk ratio, the standard error of the inverse variance summary risk ratio, 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.
- **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

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 (\texttt{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

\texttt{epi.dsl, epi.mh, epi.smd}

Examples

```r
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

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

dat          an object of class table with 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 (Brt 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.

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

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.
Note

\begin{tabular}{ccc}
\text{Obs1 +} & \text{Obs1 -} & \text{Total} \\
\hline
\text{Obs 2 +} & a & b & a+b \\
\text{Obs 2 -} & c & d & c+d \\
\text{Total} & a+c & b+d & a+b+c+d=N \\
\hline
\end{tabular}

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 \(((a + b)/N - (a + c)/N)/\). 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 <- as.table(matrix(c(19, 10, 6, 256), nrow = 2, byrow = TRUE))
colnames(dat) <- c("L1-pos", "L1-neg")
rownames(dat) <- c("L2-pos", "L2-neg")
epi.kappa(dat, method = "fleiss", alternative = "greater", conf.level = 0.95)

# 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 agreements 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

dat <- as.table(matrix(c(596, 61, 29, 987), nrow = 2, byrow = TRUE))
colnames(dat) <- c("US-pos", "US-neg")
rownames(dat) <- c("ELISA-pos", "ELISA-neg")
epi.kappa(dat, 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

```r
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 incomplete), 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,  
- fltd fat yield (kilograms) to last herd test or dry off date (computed on the basis of lactation length,  
- pltd 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

```r
## 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, 22, 18, 12, 20, 32, 27, 24, 20, 14)
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 <- data.frame(ckey, pkey, lact, dim, llen, vol, fat, pro)

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

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

epi.mh

Fixed-effects meta-analysis of binary outcomes using the Mantel-Haenszel method

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, the standard error of 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, the standard error of 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, the standard error of the Mantel-Haenszel summary odds ratio, the lower and upper bounds of the confidence interval of the Mantel-Haenszel summary odds ratio.
- RR.summary: the Mantel-Haenszel summary risk ratio, the standard error of the Mantel-Haenszel summary risk ratio, 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

data(epi.epidural)
eepi.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

Computes 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 (respec-
                    tively) 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                   the likelihood ratio of a positive and negative test.
prob                 the post-test probability of the outcome given a positive and negative test.

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?

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

## The post-test probability that this dog is hypothyroid is 24%.

## 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.

pre.pos <- 3 / (3 * 52)
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 0.2%.

---

epi.noninfb

Estimate the sample size for a parallel non-inferiority trial, binary outcomes

**Description**

Computes the sample size for a parallel non-inferiority trial with a binary outcome variable.

**Usage**

epi.noninfb(treat, control, delta, n, r = 1, power, alpha)

**Arguments**

- **treat**: the expected proportion of successes in the treatment group.
control

delta

n

r

power

alpha

Value

A list containing one or more of the following:

n.treat

n.control

n.total

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: Ps and Pn.

With a superiority trial we specify the maximum acceptable difference between Pn and Ps 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 Pn and Ps 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 Pn and Ps 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.equivb}.

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:
## Suppose it is of interest to establish non-inferiority of a new treatment
## as compared to the currently used standard treatment. A difference of less
## than 10% is of no clinical importance. Thus, the non-inferiority margin
## (delta) is set at 0.10. Assume the true cure rate for the new treatment
## is 0.85 and the control is 0.65. Assuming a one-sided test size of 2.5% and
## a power of 90% how many subjects should be included in the trial?

epi.noninfb(treat = 0.85, control = 0.65, delta = 0.10, n = NA, r = 1,
            power = 0.80, alpha = 0.025)

## A total of 558 subjects need to be enrolled in the trial, 279 in the
## treatment group and 279 in the control group.

## EXAMPLE 1 (cont.):
## Suppose only 400 subjects were enrolled in the trial, 200 in the treatment
## group and 200 in the control group. What is the estimated study power?

epi.noninfb(treat = 0.85, control = 0.65, delta = 0.10, n = 400, r = 1,
            power = NA, alpha = 0.025)

## With only 500 subjects the estimated study power is 0.66.
```

epi.noninfb Estimate the sample size for a parallel equivalence trial, continuous outcomes
Description

Computes the sample size for a parallel equivalence trial with a continuous outcome variable.

Usage

epi.noninfc(treat, control, sd, delta, n, r = 1, power, 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 a proportion.
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.
alpha scalar, defining the desired alpha level.

Value

A list containing one or more of the following:
n.treat the required number of study subject in the treatment group.
n.control the required number of study subject in the control group.
n.total the total number of study subjects required.

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: Ps and Pn.

With a superiority trial we specify the maximum acceptable difference between Pn and Ps as delta. The null hypothesis is H0: Pn - Ps <= delta and the alternative hypothesis is H1: Pn - Ps > 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 Pn and Ps as delta. The null hypothesis is H0: |Ps - Pn| >= delta and the alternative hypothesis is H1: |Ps - Pn| < 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 Pn and Ps as delta. The null hypothesis is H0: Ps - Pn >= delta and the alternative hypothesis is H1: Ps - Pn < 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.equivb.

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 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 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 of LDL is
## considered of clinically meaningful difference. Assume the standard
## 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.

epi.noninfctreat = 0.20, control = 0.20, sd = 0.10, delta = 0.05, n = NA,
               r = 1, power = 0.80, 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.
```
## 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)
```

### 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 ($\rho^c$),
- `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 ($\rho_{jk}^c$),
  - `prec`: pairwise precision values ($\rho_{jk}$),
  - `accu`: pairwise accuracy values ($\chi_{jk}^a$),
– 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


See Also

epi.ccc

Examples

```r
## Generate some artificial ratings data:
set.seed(1234)
p <- runif(10, 0, 1)
x <- replicate(n = 5, expr = rbinom(10, 4, p) + 1)

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

epi.offset

Create offset vector

Description

Creates an offset vector based on a list.

Usage

```r
epi.offset(id.names)
```

Arguments

id.names a list identifying the [location] of each case. This must be a factor.
epi.pooled

Details

This function is useful for supplying spatial data to WinBUGS.

Value

A vector of length (1 + length of 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

dat <- c(1, 1, 2, 2, 2, 3, 3, 3)
dat <- as.factor(dat)

offset <- epi.offset(dat)
offset
## [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

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

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>se</td>
<td>a vector of length one defining the sensitivity of the individual test used.</td>
</tr>
<tr>
<td>sp</td>
<td>a vector of length one defining the specificity of the individual test used.</td>
</tr>
<tr>
<td>P</td>
<td>scalar, defining the estimated true prevalence.</td>
</tr>
<tr>
<td>m</td>
<td>scalar, defining the number of individual samples to make up a pooled sample.</td>
</tr>
<tr>
<td>r</td>
<td>scalar, defining the number of pooled samples per group (or herd).</td>
</tr>
</tbody>
</table>
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
## 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 use 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*

**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**

```r
## 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 - 5557)
## Estimated number of untested pigs: 3240 (95% CI 2365 - 4797)
```

---

**epi.prcc**

Partial rank correlation coefficients

**Description**

Compute partial rank correlation coefficients.

**Usage**

```r
epi.prcc(dat, sided.test = 2)
```

**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.

**Details**

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: gamma the partial rank correlation coefficient between each input parameter and the outcome, test.statistic the test statistic used to determine the significance of non-zero values of gamma, and p.value the associated P-value.

Author(s)

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

References


Examples

```r
## Create a matrix of simulation results:
x1 <- data.frame(rnorm(n = 10, mean = 120, sd = 10))
x2 <- data.frame(rnorm(n = 10, mean = 80, sd = 5))
x3 <- data.frame(rnorm(n = 10, mean = 40, sd = 20))
y <- 2 + (0.5 * x1) + (0.7 * x2) + (0.2 * x3)
dat <- data.frame(cbind(X1 = x1, X2 = x2, X3 = x3, Y = y))
epi.prcc(dat, sided.test = 2)
```

---

**epi.prev**

*Estimate true prevalence*

Description

Computes 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", conf.level = 0.95)
```

Arguments

- **pos**: the number of positives.
- **tested**: the number tested.
- **se**: test sensitivity (0 - 1).
- **sp**: test specificity (0 - 1).
- **method**: a character string indicating the method to use. Options are "c-p" (Clopper-Pearson), "sterne" (Sterne), "blaker" (Blaker) and "wilson" (Wilson).
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).

References


Examples

## 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
## 23 were positive to the screening test. What is the estimated prevalence
## of brucellosis in this herd (and its 95% confidence interval)?

epi.prev(pos = 23, tested = 150, se = 0.96, sp = 0.89, method = "blaker",
          conf.level = 0.95)

## The estimated true prevalence of brucellosis in this herd is 5.1 cases per
## 100 cows (95% CI 0 -- 13 cases per 100 cows).

## 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?

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

## The estimated true prevalence of Helicobacter pylori in 1 -- 4 year olds is
## 0 cases per 100 (95% CI 0 -- 1.3 cases per 100).

---

epi.RtoBUGS  
R to WinBUGS data conversion

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

Does not check to ensure that only numbers are being produced. In particular, factor labels in a data
frame 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


table

epi.simplesize

Sample size under simple random sampling

description

Estimates the required sample size under simple random sampling.

Usage

epi.simplesize(N = 1E+06, Vsq, Py, epsilon.r, method = "mean", conf.level = 0.95)

Arguments

N scalar, representing the population size.

Vsq scalar, if method is total or mean this is the relative variance of the variable to be estimated (i.e. var/mean^2).

Py scalar, if method is proportion this is an estimate of the unknown population proportion.

epsilon.r the maximum relative difference between our estimate and the unknown population value.

method a character string indicating the method to be used. Options are total, mean, or proportion.

conf.level scalar, defining the level of confidence in the computed result.

Value

Returns an integer defining the size of the sample is required.

Note

epsilon.r defines the maximum relative difference between our estimate and the unknown population value. 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:

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?

```r
pmean <- 1500 * 0.05; pvar <- (300 * 0.05)^2
epi.simplesize(N = 20, Vsq = (pvar / pmean^2), Py = NA, epsilon.r = 0.20, method = "total", conf.level = 0.95)
```

## Three clinics need to be sampled to meet the survey requirements.

## 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.simplesize(N = 278, Vsq = 30^2 / 200^2, Py = NA, epsilon.r = 10/200, method = "mean", conf.level = 0.95)
```

## A total of 31 deer need to be sampled to meet the survey requirements.

## EXAMPLE 3:

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 that is seropositive to B. abortus. Calculate the required sample size.

```r
n.crude <- epi.simplesize(N = 1E+06, Vsq = NA, Py = 0.15, epsilon.r = 0.20, method = "proportion", conf.level = 0.95)
n.crude
```

## A total of 544 cattle need to be sampled to meet the survey requirements.
## EXAMPLE 3 (continued):
## Being seropositive to brucellosis is likely to cluster within herds.
## Otte and Gumm (1997) cite the intraclass correlation coefficient (rho) of
## Brucella abortus to be in the order of 0.09. Adjust the sample size
## estimate to account for clustering at the herd level. Assume that, on
## average, 20 animals will be sampled per herd:

## Let D equal the design effect and nbar equal the average number of
## individuals per cluster:

## rho = (D - 1) / (nbar - 1)

## Solving for D:
## D <- rho * (nbar - 1) + 1

rho <- 0.09; nbar <- 20
D <- rho * (nbar - 1) + 1

n.adj <- ceiling(n.crude * D)

## After accounting for the presence of clustering at the herd level we
## estimate that a total of 1475 cattle need to be sampled to meet
## the requirements of the survey.

---

### epi.smd

**Fixed-effect 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
```r
## 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.

names <- c("Audini", "Morse", "Lehman")
mean.trt <- c(41.4, 0.95, -4.10)
mean.ctrl <- c(42.3, 0.89, -3.80)
```
epi.stratasize

Sample size under stratified random sampling

Description

Estimates the required sample size under stratified random sampling.

Usage

epi.stratasize(strata.n, strata.mean, strata.var, strata.Py, epsilon.r, method = "mean", conf.level = 0.95)

Arguments

- **strata.n**: vector, defining the size of each strata.
- **strata.mean**: vector, representing the expected means in each strata. Only used when method = "mean", "total" or "pps".
- **strata.var**: vector, representing the expected variance in each strata. Only used when method = "mean", "total" or "pps".
- **strata.Py**: vector, representing the expected proportions in each strata. Only used when method = "proportion".
- **epsilon.r**: the maximum relative difference between our estimate and the unknown population value.
- **method**: a character string indicating the method to be used. Options are mean, total, proportion, or pps.
- **conf.level**: scalar, defining 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**: 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.

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)
Note

Use method proportion to estimate sample size using stratified random sampling with equal weights (see Levy and Lemeshow, page 176). Use method pps to estimate sample size using proportional stratified random sampling with proportional allocation (see Levy and Lemeshow, page 179).

When method = "proportion" the vectors strata.mean and strata.var are ignored.

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

```r
# EXAMPLE 1:
# Hospital episodes (Levy and Lemeshow 1999, page 176 -- 178)
# 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?

strata.n <- c(600, 500, 400)
strata.mean <- c(0.164, 0.166, 0.236)
strata.var <- c(0.245, 0.296, 0.436)
epi.stratasize(strata.n, strata.mean, strata.var, strata.Py,
epsilon.r = 0.20, method = "mean", conf.level = 0.95)

# The number allocated to the under 45 years, 45 -- 64 years, and 65 years
# and over stratus should be 223, 186, and 149 (a total of 558). These
# results differ from the worked example provided in Levy and Lemeshow where
# certainty is set to approximately 99%.

# EXAMPLE 2:
# 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
```
## Description

Computes the sample size, power, and minimum detectable difference for cohort studies (using count data), case-control studies, when comparing means and survival.

## Usage

```r
epi.studysize(treat, control, n, sigma, power, r = 1, design = 1, sided.test = 2, conf.level = 0.95, method = "means")
```

## 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`: when `method = "means"` this is the expected standard deviation of the variable of interest for both treatment and control groups. When `method = "case.control"` this is the expected proportion of study subjects exposed to the risk factor of interest. This argument is ignored when `method = "proportions",method = "survival", or method = "cohort.count".
- `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 treatment group is better or worse than the control group. Use a one-sided test to evaluate whether or not the treatment group is better than the control group.
- `conf.level`: scalar, defining the level of confidence in the computed result.
- `method`: a character string indicating the method to be used. Options are `means`, `proportions`, `survival`, `cohort.count`, or `case.control`.
Details

The methodologies adopted in this function follow closely the approach described in Chapter 8 of Woodward (2005).

When method = "means" the argument treat defines the mean outcome for the treatment group, control defines the mean outcome for the control group, and sigma defines the standard deviation of the outcome, assumed to be the same across the treatment and control groups (see Woodward pp 397 - 403).

When method = "proportions" the argument treat defines the proportion in the treatment group and control defines the proportion in the control group. The arguments sigma and r are ignored.

When method = "survival" 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. The argument sigma is ignored (see Therneau and Grambsch pp 61 - 65).

When method = "cohort.count" the argument treat defines the estimated incidence risk (cumulative incidence) of the event of interest in the treatment group and control defines the estimated incidence risk of the event of interest in the control group. The argument sigma is ignored (see Woodward pp 405 - 410).

When method = "case.control" the argument treat defines the estimated incidence risk (cumulative incidence) of the event of interest in the treatment group and control defines the estimated incidence risk of the event of interest in the control group. The argument sigma is the expected proportion of study subjects exposed to the risk factor of interest (see Woodward pp 410 - 412).

In case-control studies sample size estimates are worked out on the basis of an expected odds (or risk) ratio. When method = "case.control" the estimated incidence risk estimates in the treat and control groups are used to define the expected risk ratio. See example 7 below, taken from Woodward p 412.

For method = "proportions" 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. In contrast, method = "cohort.count" relates to the two-sample problem where neither proportion is known (or assumed, at least). Thus, there is much more uncertainty in the method = "cohort.count" situation (compared with method = "proportions") and correspondingly a requirement for a much larger sample size. Generally, method = "cohort.count" is more useful in practice. method = "proportions" is used in special 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 one or more of the following:

- **n.crude**: the crude estimated total number of subjects required for the specified level of confidence and power.
- **n.total**: the total estimated 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.
- **delta**: the minimum detectable difference given the specified level of confidence and power.
lambda the minimum detectable risk ratio >1 and the maximum detectable risk ratio <1.

power the power of the study given the specified number of study subjects and power.

Note

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

The odds ratio and the risk ratio are approximately equal when the event of interest is rare. In this function method = "case.control" returns the sample size required to detect an approximate risk ratio in a case-control study (see Woodward p 412).

When method = "proportions" values need to be entered for control, n, and power to return a value for delta. When method = "cohort.count" values need to be entered for control, n, and power to return a value for lambda (see example 6 below).

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.studysize(treat = 5, control = 4.5, n = NA, sigma = 1.4, power = 0.95,
             r = 1, design = 1, sided.test = 1, conf.level = 0.95, method = "means")
```

## To satisfy the study requirements 340 individuals need to be tested: 170 in
## the first year and 170 in the second year.

```r
## 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 wish 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 idea of its effect.
## How many women should be enrolled in this study?

```r
epi.studysize(treat = 8, control = 4, n = NA, sigma = 4, power = 0.90, 
r = 4, design = 1, sided.test = 2, conf.level = 0.95, method = "means")
```

## The estimated sample size is 66. We round this up to the nearest multiple of 5, to 70. We allocate 70/5 = 14 women to the placebo group and four times as many (56) to the iron treatment group.

## Example 3 (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?

```r
epi.studysize(treat = 0.30, control = 0.32, n = NA, sigma = NA, power = 0.90, 
r = 1, design = 1, sided.test = 1, conf.level = 0.95, method = "proportions")
```

## # # A total of 18,315 men should be sampled: 9158 in the treatment group and 9158 in the control group.

## Example 4 (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?

```r
epi.studysize(treat = 0.45, control = 0.30, n = NA, sigma = NA, power = 0.90, 
r = 1, design = 1, sided.test = 2, conf.level = 0.95, method = "survival")
```

## 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 5 (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?

```r
epi.studysize(treat = NA, control = NA, n = 500, sigma = NA, power = 0.90, 
r = 1, design = 1, sided.test = 2, conf.level = 0.95, method = "survival")
```
## 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.

## EXAMPLE 6 (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 rate 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?

treat = 1.4 * (5 * 413)/100000
control = (5 * 413)/100000
epi.studysize(treat = treat, control = control, n = NA, sigma = NA,
              power = 0.90, r = 1, design = 1, sided.test = 1, conf.level = 0.95,
              method = "cohort.count")

## A total of 12,130 men need to be sampled (6065 smokers and 6065 non-smokers).

## EXAMPLE 7 (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?

control = (5 * 413)/100000
epi.studysize(treat = NA, control = control, n = 5000, sigma = NA,
              power = 0.90, r = 1, design = 1, sided.test = 1, conf.level = 0.95,
              method = "cohort.count")

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

## EXAMPLE 8 (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 are needed to detect an approximate risk
## ratio of 2.0 with 0.90 power using a two-sided 0.05 test? Previous surveys
## have shown that around 0.30 of the male population are smokers.

epi.studysize(treat = 2/100, control = 1/100, n = NA, sigma = 0.30,
              power = 0.90, r = 1, design = 1, sided.test = 2, conf.level = 0.95,
              method = "case.control")

## A total of 376 men need to be sampled: 188 cases and 188 controls.
## EXAMPLE 9 (from Woodward p 414):

Suppose we wish to determine the power to detect an approximate risk ratio of 2.0 using a two-sided 0.05 test when 188 cases and 940 controls are available (that is, the ratio of cases to controls is 1:5). Assume the prevalence of smoking in the male population is 0.30.

```r
n <- 188 + 940
epi.studysize(treat = 2/100, control = 1/100, n = n, sigma = 0.30,
              power = NA, r = 0.2, design = 1, sided.test = 2, conf.level = 0.95,
              method = "case.control")
```

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

## EXAMPLE 10:

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 group to be 0.30 and the proportion of cows responding in the control group to be 0.15? The required power for this study is 0.80 using a two-sided 0.05 test.

```r
epi.studysize(treat = 0.30, control = 0.15, n = NA, sigma = NA,
              power = 0.80, r = 1, design = 1, sided.test = 2, conf.level = 0.95,
              method = "cohort.count")
```

A total of 242 cows are required: 121 in the treatment group and 121 in the control 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 across herds.

For the exercise, assume that the intra-cluster correlation coefficient (the rate of homogeneity, rho) is 0.05 and the average number of cows per herd is 30. Calculate the design effect, given 

\[
\text{rho} = (\text{design} - 1) / (\text{nbar} - 1),
\]

where nbar equals the average number of individuals per cluster:

```r
design <- 0.05 * (30 - 1) + 1
epi.studysize(treat = 0.30, control = 0.15, n = NA, sigma = NA,
              power = 0.80, r = 1, design = design, sided.test = 2, conf.level = 0.95,
              method = "cohort.count")
```

A total of 592 cows are required for this study: 296 in the treatment group and 296 in the control group.
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
```r
epi.tests(dat, conf.level = 0.95)
```

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

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

Arguments
dat an object of class table 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:
aprev apparent prevalence.
tprev true prevalence.
se test sensitivity.
sp test specificity.
diag.acc diagnostic accuracy.
diag.or diagnostic odds ratio.
nnd number needed to diagnose.
youden
ppv
npv
plr
nlr

Youden’s index.
positive predictive value.
negative predictive value.
likelihood ratio of a positive test.
likelihood ratio of a negative test.

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).

References


Examples

```r
## Scott et al. 2008, Table 1:
## A new diagnostic test was trialled on 1586 patients. Of 744 patients
## that were disease positive, 670 tested positive. Of 842 patients that
## were disease negative, 640 tested negative. What is the likelihood
## ratio of a positive test? What is the number needed to diagnose?
```
dat <- as.table(matrix(c(670,202,74,640), nrow = 2, byrow = TRUE))
colnames(dat) <- c("Dis+","Dis-")
rownames(dat) <- c("Test+","Test-")
rval <- epi.tests(dat, conf.level = 0.95)
print(rval); summary(rval)

## Test sensitivity is 0.90 (95% CI 0.88 -- 0.92). Test specificity is
## 0.76 (95% CI 0.73 -- 0.79). The likelihood ratio of a positive test
## is 3.75 (95% CI 3.32 to 4.24). 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.
Index

*Topic **datasets**
  - epi.epidural, 47
  - epi.incin, 53
  - epi.SC1ip, 83

*Topic **htest**
  - epi.occc, 75

*Topic **univar**
  - epi.2by2, 3
  - epi.about, 12
  - epi.asc, 13
  - epi.betabuster, 14
  - epi.bohning, 16
  - epi.ccc, 17
  - epi.cluster1size, 19
  - epi.cluster2size, 20
  - epi.clustersize, 25
  - epi.conf, 27
  - epi.convgrid, 32
  - epi.cp, 33
  - epi.cpresids, 34
  - epi.descriptives, 35
  - epi.detectsize, 36
  - epi.dgamma, 38
  - epi.directadj, 39
  - epi.dms, 41
  - epi.dsl, 42
  - epi.edr, 44
  - epi.empbayes, 45
  - epi.equivb, 47
  - epi.equivc, 50
  - epi.herdttest, 52
  - epi.indirectadj, 54
  - epi.insthaz, 57
  - epi.interaction, 58
  - epi.iv, 60
  - epi.kappa, 62
  - epi.ltd, 66
  - epi.mh, 67
  - epi.nomogram, 69
  - epi.noninfb, 70
  - epi.noninfrc, 72
  - epi.offset, 76
  - epi.pooled, 77
  - epi.popsize, 78
  - epi.prcc, 79
  - epi.prev, 80
  - epi.RtoBUGS, 82
  - epi.simplesize, 84
  - epi.smd, 86
  - epi.stratasize, 88
  - epi.studysize, 90
  - epi.tests, 95
  - epi.2by2, 3
  - epi.about, 12
  - epi.asc, 13
  - epi.betabuster, 14
  - epi.bohning, 16
  - epi.ccc, 17, 76
  - epi.cluster1size, 19
  - epi.cluster2size, 20
  - epi.clustersize, 25
  - epi.conf, 27
  - epi.convgrid, 32
  - epi.cp, 33, 34, 35
  - epi.cpresids, 34
  - epi.descriptives, 35
  - epi.detectsize, 36
  - epi.dgamma, 38
  - epi.directadj, 39, 55
  - epi.dms, 41
  - epi.dsl, 42, 62, 69, 87
  - epi.edr, 44
  - epi.empbayes, 45
  - epi.epidural, 47
  - epi.equivb, 47, 51, 71, 74
  - epi.equivc, 50
  - epi.herdttest, 52
  - epi.incin, 53
epi.indirectadj, 40, 54
epi.insthaz, 57
epi.interaction, 58
epi.iv, 44, 60, 69, 87
epi.kappa, 62
epi.ltd, 66
epi.mh, 44, 62, 67, 87
epi.nomogram, 69
epi.noninfb, 70
epi.noninfc, 72
epi.occc, 18, 75
epi.offset, 76
epi.pooled, 77
epi.popsiz, 78
epi.prcc, 79
epi.prev, 80
epi.RtoBUGS, 82
epi.SClip, 83
epi.simplesize, 84
epi.smd, 44, 62, 69, 86
epi.stratasize, 88
epi.studysize, 90
epi.tests, 95

print.epi.2by2 (epi.2by2), 3
print.epi.occc (epi.occc), 75
print.epi.tests (epi.tests), 95

summary.epi.2by2 (epi.2by2), 3
summary.epi.occc (epi.occc), 75
summary.epi.tests (epi.tests), 95