

Package ‘CPTtools’

July 28, 2015

Version 0.3-2

Date 2015/06/29

Title Tools for creating conditional probability tables

Author Russell Almond

Maintainer Russell Almond <ralmond@fsu.edu>

Depends R (>= 3.0)

Description Provides support for parameterized tables for Bayesian networks .

License Artistic-2.0

URL <http://pluto.coe.fsu.edu/RNetica>

R topics documented:

CPTtools-package	2
ACED.scores	8
areaProbs	12
betaci	14
buildFactorTab	15
buildParentList	17
buildRegressions	18
buildRegressionTables	19
calcDDTable	21
calcDNTable	23
calcDPCTable	26
calcDSlike	30
calcDSTable	32
calcNoisyAndTable	35
calcNoisyOrTable	37
colspread	40
compareBars	41
Compensatory	43
dataTable	45
effectiveThetas	46
eThetaFrame	47

getTableStates	49
gradedResponse	50
localDepTest	52
mapDPC	54
MathGrades	56
mcSearch	58
mutualInformation	59
numericPart	61
OCP	62
OffsetConjunctive	64
parseProbVec	66
partialCredit	67
proflevelci	69
readHistory	71
rescaleTable	72
scaleMatrix	73
scaleTable	74
stackedBarplot	75
stackedBars	78
structMatrix	79
woeBal	80
woeHist	82
Index	85

CPTtools-package	<i>Tools for creating conditional probability tables</i>
------------------	--

Description

Provides support for parameterized tables for Bayesian networks .

Details

The DESCRIPTION file: This package was not yet installed at build time.

CPTtools is a collection of various bits of R code useful for processing Bayes net output. Some were designed to work with ETS's proprietary StatShop code, and some with RNetica. The code collected in this package is all free from explicit dependencies on the specific Bayes net package and will hopefully be useful with other systems as well.

The majority of the code are related to building conditional probability tables (CPTs) for Bayesian networks. The package has two output representations for a CPT. The first is a `data.frame` object where the first several columns are factor variables corresponding the the parent variables, and the remaining columns are numeric variables corresponding to the state of the child variables. The rows represent possible configurations of the parent variables. An example is shown below.

	S1	S2	Full	Partial	None
1	High	High	0.81940043	0.15821522	0.02238436
2	Medium	High	0.46696668	0.46696668	0.06606664
3	Low	High	0.14468106	0.74930671	0.10601223
4	High	Medium	0.76603829	0.14791170	0.08605000
5	Medium	Medium	0.38733177	0.38733177	0.22533647
6	Low	Medium	0.10879020	0.56342707	0.32778273
7	High	Low	0.65574465	0.12661548	0.21763987
8	Medium	Low	0.26889642	0.26889642	0.46220715
9	Low	Low	0.06630741	0.34340770	0.59028489
10	High	LowerYet	0.39095414	0.07548799	0.53355787
11	Medium	LowerYet	0.11027649	0.11027649	0.77944702
12	Low	LowerYet	0.02337270	0.12104775	0.85557955

The second representation is a table (matrix) with just the numeric part. Two approaches to building these tables from parameters are described below. The more flexible discrete partial credit model is used for the basis of the parameterized networks in the [Peanut](#) package.

In addition to the code for building partial credit networks, this package contains some code for building Bayesian network structures from (inverse) correlation matrixes, and graphical displays for Bayes net output. The latter includes some diagnostic plots and additional diagnostic tests.

Discrete Partial Credit Framework

The original parameterization for creating conditional probability tables based on Almond et al (2001) proved to be insufficiently flexible. Almond (2015) describes a newer parameterization based on three steps:

1. Translate the parent variables onto a numeric effective theta scale ([effectiveThetas](#)).
2. Combine the parent effective thetas into a single effective theta using a combination rule ([Compensatory](#), [OffsetConjunctive](#)).
3. Convert the effective theta for each row of the table into conditional probabilities using a link function ([gradedResponse](#), [partialCredit](#)).

The [partialCredit](#) link function is particularly flexible as it allows different parameterizations and different combination rules for each state of the child variable. This functionality is best captured by the two high level functions:

[calcDPCTable](#) Creates the probability table for the discrete partial credit model given the parameters.

[mapDPC](#) Finds an MAP estimate for the parameters given an observed table of counts.

This parameterization serves as basis for the model used in the [Peanut](#) package.

Other parametric CPT models

The first two steps of the discrete partial credit framework outlined above are due to a suggestion by Lou DiBello (Almond et al, 2001). This lead to an older framework, in which the link function was hard coded into the conditional probability table formation. The models were called DiBello-XX, where XX is the name of the link function. Almond et al. (2015) describes several additional examples.

`calcDDTable` Calculates DiBello-Dirichlet model probability and parameter tables.
`calcDNTable` Creates the probability table for DiBello-Normal distribution.
`calcDSTable` Creates the probability table for DiBello-Samejima distribution.
`calcDSlike` Calculates the log-likelihood for data from a DiBello-Samejima (Normal) distribution.

Diez (1993) and Srinivas (1993) describe an older parametric framework for Bayes nets based on the noisy-or or noisy-max function. These are also available.

`calcNoisyAndTable` Calculate the conditional probability table for a Noisy-And or Noisy-Min distribution.

`calcNoisyOrTable` Calculate the conditional probability table for a Noisy-Or distribution.

Building Bayes nets from (inverse) correlation matrixes

Almond (2010) noted that in many cases the best information about the relationship among variables came from a procedure that produces a correlation matrix (e.g., a factor analysis). Applying a trick from Whittaker (1990), connecting pairs of nodes corresponding to nonzero entries in an inverse correlation matrix produces an undirected graphical model. Ordering in the nodes in a perfect ordering allows the undirected model to be converted into a directed model (Bayesian network). The conditional probability tables can then be created through a series of regressions.

The following functions implement this protocol:

`structMatrix` Finds graphical structure from a covariance matrix.
`mcSearch` Orders variables using Maximum Cardinality search.
`buildParentList` Builds a list of parents of nodes in a graph.
`buildRegressions` Creates a series of regressions from a covariance matrix.
`buildRegressionTables` Builds conditional probability tables from regressions.

Other model construction tools

These functions are a grab bag of lower level utilities useful for building CPTs:

`areaProbs` Translates between normal and categorical probabilities.
`numericPart` Splits a mixed data frame into a numeric matrix and a factor part..
`dataTable` Constructs a table of counts from a setof discrete observations..
`eThetaFrame` Constructs a data frame showing the effective thetas for each parent combination..
`effectiveThetas` Assigns effective theta levels for categorical variable.
`getTableStates` Gets meta data about a conditional probability table..
`rescaleTable` Rescales the numeric part of the table.
`scaleMatrix` Scales a matrix to have a unit diagonal.
`scaleTable` Scales a table according to the Sum and Scale column.

Bayes net output displays and tests

Almond et al. (2009) suggested using hanging barplots for displaying Bayes net output and gives several examples. The function `stackedBars` produces the simple version of this plot and the function `compareBars` compares two distributions (e.g., prior and posterior). The function `buildFactorTab` is useful for building the data and the function `colorspread` is useful for building color gradients.

Madigan, Mosurski and Almond (1997) describe a graphical weight of evidence balance sheet (see also Almond et al, 2015, Chapter 7; Almond et al, 2013). The function `woeHist` calculates the weights of evidence for a series of observations and the function `woeBal` produces a graphical display.

Sinharay and Almond (2006) propose a graphical fit test for conditional probability tables (see also, Almond et al, 2015, Chapter 10). The function `OCP` implements this test, and the function `betaci` creates the beta credibility intervals around which the function is built.

The key to Bayesian network models are the assumptions of conditional independence which underlie the model. The function `localDepTest` tests these assumptions based on observed (or imputed) data tables.

The function `mutualInformation` calculates the mutual information of a two-way table, a measure of the strength of association. This is similar to the measure used in many Bayes net packages (e.g., `MutualInfo`).

Data sets

Two data sets are provided with this package:

`ACED` Data from ACED field trial (Shute, Hansen, and Almond, 2008). This example is based on a field trial of a Bayesian network based Assessment for Learning system, and contains both item-level response and high-level network summaries. A complete description of the Bayes net can be found at <http://ecd.ralmond.net/ecdwiki/ACED/ACED>.

`MathGrades` Grades on 5 mathematics tests from Mardia, Kent and Bibby (from Whittaker, 1990).

Index

Complete index of all functions.

Index: This package was not yet installed at build time.

Author(s)

Russell Almond

Maintainer: Russell Almond <ralmond@fsu.edu>

References

Almond, R.G. (2015). An IRT-based Parameterization for Conditional Probability Tables. Paper submitted to the 2015 Bayesian Application Workshop at the Uncertainty in Artificial Intelligence conference.

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (2015) *Bayesian Networks in Educational Assessment*. Springer.

- Almond, R. G. (2010). 'I can name that Bayesian network in two matrixes.' *International Journal of Approximate Reasoning*, **51**, 167-178.
- Almond, R. G., Shute, V. J., Underwood, J. S., and Zapata-Rivera, J.-D (2009). Bayesian Networks: A Teacher's View. *International Journal of Approximate Reasoning*, **50**, 450-460.
- Almond, R.G., DiBello, L., Jenkins, F., Mislevy, R.J., Senturk, D., Steinberg, L.S. and Yan, D. (2001) Models for Conditional Probability Tables in Educational Assessment. *Artificial Intelligence and Statistics 2001* Jaakkola and Richardson (eds), Morgan Kaufmann, 137-143.
- Diez, F. J. (1993) Parameter adjustment in Bayes networks. The generalized noisy OR-gate. In Heckerman and Mamdani (eds) *Uncertainty in Artificial Intelligence 93*. Morgan Kaufmann. 99-105.
- Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, **16**, 159-176. DOI: 10.1177/014662169201600206
- Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, **34**, (No. 4, Part 2).
- Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it—Or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence and Education*, **18**(4), 289-316.
- Sinharay, S. and Almond, R.G. (2006). Assessing Fit of Cognitively Diagnostic Models: A case study. *Educational and Psychological Measurement*, **67**(2), 239-257.
- Srinivas, S. (1993) A generalization of the Noisy-Or model, the generalized noisy OR-gate. In Heckerman and Mamdani (eds) *Uncertainty in Artificial Intelligence 93*. Morgan Kaufmann. 208-215.
- Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.
- Madigan, D., Mosurski, K. and Almond, R. (1997) Graphical explanation in belief networks. *Journal of Computational Graphics and Statistics*, **6**, 160-181.
- Almond, R. G., Kim, Y. J., Shute, V. J. and Ventura, M. (2013). Debugging the Evidence Chain. In Almond, R. G. and Mengshoel, O. (Eds.) *Proceedings of the 2013 UAI Application Workshops: Big Data meet Complex Models and Models for Spatial, Temporal and Network Data (UAI2013AW)*, 1-10. <http://ceur-ws.org/Vol-1024/paper-01.pdf>

See Also

[RNetica](#) ~~ [Peanut](#) ~~

Examples

```
## Set up variables
skill1l <- c("High", "Medium", "Low")
skill2l <- c("High", "Medium", "Low", "LowerYet")
correctL <- c("Correct", "Incorrect")
pcreditL <- c("Full", "Partial", "None")
gradeL <- c("A", "B", "C", "D", "E")

## New Discrete Partial Credit framework:

## Complex model, different rules for different levels
```

```

cptPC2 <- calcDPCFrame(list(S1=skill11l,S2=skill12l),pcreditL,
  list(full=log(1),partial=log(c(S1=1,S2=.75))),
  betas=list(full=c(0,999),partial=1.0),
  rule=list("OffsetDisjunctive","Compensatory"))

## Graded Response using the older DiBello-Samejima framework.
cptGraded <- calcDSTable(list(S1=skill11l),gradeL, 0.0, 0.0, dinc=c(.3,.4,.3))

## Building a Bayes net from a correlation matrix.
data(MathGrades)
pl <- buildParentList(structMatrix(MathGrades$var),"Algebra")
rt <- buildRegressions(MathGrades$var,MathGrades$means,pl)
tabs <- buildRegressionTables(rt, MathGrades$pvecs, MathGrades$means,
  sqrt(diag(MathGrades$var)))

## Stacked Barplots:
margins.prior <- data.frame (
  Trouble=c(Novice=.19,Semester1=.24,Semester2=.28,Semester3=.20,Semester4=.09),
  NDK=c(Novice=.01,Semester1=.09,Semester2=.35,Semester3=.41,Semester4=.14),
  Model=c(Novice=.19,Semester1=.28,Semester2=.31,Semester3=.18,Semester4=.04)
)

margins.post <- data.frame(
  Trouble=c(Novice=.03,Semester1=.15,Semester2=.39,Semester3=.32,Semester4=.11),
  NDK=c(Novice=.00,Semester1=.03,Semester2=.28,Semester3=.52,Semester4=.17),
  Model=c(Novice=.10,Semester1=.25,Semester2=.37,Semester3=.23,Semester4=.05))

stackedBars(margins.post,3,
  main="Marginal Distributions for NetPASS skills",
  sub="Baseline at 3rd Semester level.",
  cex.names=.75, col=HSV(223/360,.2,0.10*(5:1)+.5))

compareBars(margins.prior,margins.post,3,c("Prior","Post"),
  main="Margins before/after Medium Trouble Shooting Task",
  sub="Observables: cfgCor=Medium, logCor=High, logEff=Medium",
  legend.loc = "topright",
  cex.names=.75, col1=HSV(h=.1,s=.2*1:5-.1,alpha=1),
  col2=HSV(h=.6,s=.2*1:5-.1,alpha=1))

## Weight of evidence balance sheets
sampleSequence <- read.csv(paste(library(help="CPTtools")$path,
  "testFiles","SampleStudent.csv",
  sep=.Platform$file.sep),
  header=TRUE,row.names=1)

woeBal(sampleSequence[,c("H","M","L")],c("H"),c("M","L"),lcex=1.25)

### Observable Characteristic Plot
pi <- c("+=".15,"-=".85)
nnn <- c("(0,0,0)"=20,"(0,0,1)"=10,
  "(0,1,0)"=10,"(0,1,0)"=5,
  "(1,0,0)"=10,"(1,0,1)"=10,

```

```

      "(1,1,1)"=10,"(1,1,1)"=25)
xx1 <- c("(0,0,0)"=2,"(0,0,1)"=5,
      "(0,1,0)"=1,"(0,1,1)"=3,
      "(1,0,0)"=0,"(1,0,1)"=2,
      "(1,1,0)"=5,"(1,1,1)"=24)
grouplabs <- c(rep("-",3),"+")
grouplabs1 <- rep(grouplabs,each=2)
OCP2 (xx1,nnn,grouplabs1,pi,c("-", "+"),ylim=c(0,1), reflty=c(2,4),
      setlabs=c("Low Skill3","High Skill3"),setat=-.8,
      main="Data for which Skill 3 is relevant")

```

ACED.scores

Data from ACED field trial

Description

ACED (Adaptive Content with Evidence-Based Diagnosis; Shute, Hansen and Almond, 2008) is a Bayes net based assessment system which featured: (a) adaptive item selection and (b) extended feedback for incorrect items. This data contains both item level and pretest/posttest data from a field trial of the ACED system.

Usage

```
data("ACED")
```

Format

ACED contains 3 data.frame objects and one explanatory variable.

ACED.scores is data frame with 230 observations on 74 variables. These are mostly high-level scores from the Bayesian network.

Cond_code a numeric vector giving the experimental condition for this student, see also Cond

Seq a factor describing whether the sequence of items was Linear or Adaptive

FB a factor describing whether the feedback for incorrect items was Extended or AccuracyOnly

All_Items a numeric vector giving the number of items in ACED

Correct a numeric vector giving the number of items the student got correct

Incorr a numeric vector giving the number of items the student got incorrect

Remain a numeric vector giving the number of items not reached or skipped

ElapTime a numeric vector giving the total time spent on ACED The next group of columns give "scores" for each of the nodes in the Bayesian network. Each node has four scores, and the columns are names *pnodeScoreType* where *node* is replaced by one of the codes in ACED.allskills.

pnodeH a numeric vector giving the probability *node* is in the high state

pnodeM a numeric vector giving the probability *node* is in the medium state

pnodeL a numeric vector giving the probability *node* is in the low state

EAPnode the expected a posteriori value of *node* assuming an equal interval scale, where L=1, M=2 and H=3

MAPnode a factor vector giving maximum a posteriori value of *node*, i.e., which.max(*pnodeH*, *pnodeM*, *pnodeL*). After a number of columns with this pattern, the last column is:

Cond a factor describing the experimental condition with levels Adaptive/Accuracy, Adaptive/Extended and Linear/Extended

ACED.skillNames is a character vector giving the abbreviations used for the node names. Here are the interpretations:

sgp Solve Geometric Problems. This is the highest level variable for the field trial data.

arg Algebraic Rule Geometric

cr Find Common Ratio

dt Distinguish Types of series

exa Examples (Geometric)

exp Explicit Rule (Geometric)

ext Extend Series (Geometric)

ind Induce Rules (Geometric)

mod Model (Geometric)

rec Recursive Rules (Geometric)

tab Tabular Representations (Geometric)

ver Verbal Rules (Geometric)

pic Pictorial Representations (Geometric)

ACED.scores is data frame with 230 observations on 73 variables. These are mostly item-level scores from the field trial.

Cond_code a numeric vector giving the experimental condition for this student, see also **Cond**

Seq a factor describing whether the sequence of items was Linear or Adaptive

FB a factor describing whether the feedback for incorrect items was Extended or AccuracyOnly

All_Items a numeric vector giving the number of items in ACED

Correct a numeric vector giving the number of items the student got correct

Incorr a numeric vector giving the number of items the student got incorrect

Remain a numeric vector giving the number of items not reached or skipped

ElapTime a numeric vector giving the total time spent on ACED The next 63 columns represent the items from the ACED assessment. All are factor variables, with possible values Incorrect and Correct. The variables are named all named t (for task) followed by the name of one or more variables tapped by the task (if there is more than one, then the first one is “primary”.) This is followed by a numeric code, 1, 2 or 3, giving the difficulty (easy, medium or hard) and a letter (a, b or c) used to indicate alternate tasks following the same task model. Finally, following a period, there is a version number (all of the tasks are version 1).

After the variables, the last column is:

Cond a factor describing the experimental condition with levels Adaptive/Accuracy, Adaptive/Extended and Linear/Extended

ACED.prePost is data frame with 290 observations on 32 variables giving the results of the pretest and posttest.

Cond_code a numeric vector giving the experimental condition for this student, see also Cond

Seq a factor describing whether the sequence of items was Linear or Adaptive

FB a factor describing whether the feedback for incorrect items was Extended or AccuracyOnly

All_Items a numeric vector giving the number of items in ACED

Form_Order a factor variables describing whether (AB) Form A was the pretest and Form B was the posttest or (BA) vise versa.

Level_Code a factor variable describing the academic track of the student with levels Honors, Academic, Regular, Part 1, Part 2 and ELL. The codes Part 1 and Part 2 refer to special education students in Part 1 (mainstream classroom) or Part 2 (sequestered).

PreACorr corrected score on Form A for students who took Form A as a pretest

PostBCorr corrected score on Form B for students who took Form B as a posttest

PreBCorr corrected score on Form B for students who took Form B as a pretest

PostACorr corrected score on Form A for students who took Form A as a posttest

PreScore a numeric vector with either the non-missing value from PreACorr and PreBCorr

PostScore a numeric vector with either the non-missing value from PostACorr and PostBCorr

Gender a factor variable giving the (self-reported) gender of the student (codebook is lost)

Race a factor variable giving the (self-reported) race of the student (codebook is lost)

Gain PostScore - PreScore

preacorr_adj PreACorr adjusted to put forms A and B on the same scale

postbcorr_adj PostBCorr adjusted to put forms A and B on the same scale

prebcorr_adj PreBCorr adjusted to put forms A and B on the same scale

postacorr_adj PostACorr adjusted to put forms A and B on the same scale

Zpreacorr_adj standardized version of preacorr_adj

Zpostbcorr_adj standardized version of postbcorr_adj

Zprebcorr_adj standardized version of prebcorr_adj

Zpostacorr_adj standardized version of postacorr_adj

scale_prea score on Form A for students who took Form A as a pretest scaled to range 0-100

scale_preb score on Form B for students who took Form B as a pretest scaled to range 0-100

pre_scaled scale score on pretest (whichever form)

scale_posta score on Form A for students who took Form A as a posttest scaled to range 0-100

scale_postb score on Form B for students who took Form B as a posttest scaled to range 0-100

post_scaled scale score on pretest (whichever form)

gain_scaled post_scaled - pre_scaled

Flagged a logical variable (codebook lost)

Cond a factor describing the experimental condition with levels Adaptive/Accuracy, Adaptive/Extended, Linear/Extended and Control

Details

ACED is a Bayesian network based Assessment for Learning learning system, thus it served as both a assessment and a tutoring system. It had two novel features which could be turned on and off, elaborated feedback (turned off, it provided accuracy only feedback) and adaptive sequencing of items (turned off, it scheduled items in a fixed linear sequence).

It was originally built to cover all algebraic sequences (arithmetic, geometric and other recursive), but only the branch of the system using geometric sequences was tested. Shute, Hansen and Almond (2008) describe the field trial. Students from a local middle school (who studied arithmetic, but not geometric sequences as part of their algebra curriculum) were recruited for the study. The students were randomized into one of four groups:

Adaptive/Accuracy Adaptive sequencing was used, but students only received correct/incorrect feedback.

Adaptive/Extended Adaptive sequencing was used, but students received extended feedback for incorrect items.

Linear/Extended The fixed linear sequencing was used, but students received extended feedback for incorrect items.

Control The students did independent study and did not use ACED.

Because students in the control group were not exposed to the ACED task, neither the Bayes net level scores nor the item level scores are available for those groups, and those students are excluded from ACED.scores and ACED.items. The students are in the same order in all of the data sets, with the 60 control students tacked onto the end of the ACED.prePost data set.

All of the students (including the control students) were given a 25-item pretest and a 25-item posttest with items similar to the ones used in ACED. The design was counterbalanced, with half of the students receiving Form A as the pretest and Form B as the posttest and the other half the other way around, to allow the two forms to be equated using the pretest data. The details are buried in ACED.prePost.

Note that some irregularities were observed with the English Language Learner (ACED.prePost\$Level_code=="ELL") students. Their teachers were allowed to translated words for the students, but in many cases actually wound up giving instruction as part of the translation.

Source

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it—Or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence and Education*, **18**(4), 289-316.

Thanks to Val Shute for permission to use the data.

ACED development and data collection was sponsored by National Science Foundation Grant No. 0313202.

References

A more detailed description, including a Q-matrix can be found at the ECD Wiki: <http://ecd.ralmond.net/ecdwiki/ACED/ACED>.

Examples

```
data(ACED)
```

areaProbs

Translates between normal and categorical probabilities

Description

Maps between a continuous (normal) variable and a discrete variable by establishing a set of bins to maintain a particular probability vector. The `pvecToCutpoints` function returns the cut points separating the bins, the `pvecToMidpoints` returns a central point from each bin, and the `areaProbs` calculates the fraction of a normal curve falling into a particular bin.

Usage

```
pvecToCutpoints(pvec, mean = 0, std = 1)
pvecToMidpoints(pvec, mean = 0, std = 1)
areaProbs(pvec, condmean, condstd, mean = 0, std = 1)
```

Arguments

<code>pvec</code>	A vector of marginal probabilities for the categories of the discrete variable. Elements should be ordered from smallest to largest.
<code>mean</code>	The mean of the continuous variable.
<code>std</code>	The standard deviation of the continuous variable.
<code>condmean</code>	The conditional mean of the continuous variable.
<code>condstd</code>	The conditional standard deviation of the continuous variable.

Details

Let S be a discrete variable whose states s_k are given by `names(pvec)[k]` and for which the marginal probability $Pr(S = s_k) = p_k$ is given by `pvec[k]`. Let Y be a continuous normal variable with mean `mean` and standard deviation `std`. These function map between S and Y .

The function `pvecToCutpoints` produces a series of cutpoints, c_k , such that setting s_k to S when $c_k \leq Y \leq c_{k+1}$ produces the marginal probability specified by `pvec`. Note that c_1 is always `-Inf` and c_{K+1} is always `Inf` (where K is `length(pvec)`).

The function `pvecToMidpoints` produces the midpoints (with respect to the normal density) of the intervals defined by `pvecToCutpoints`. In particular, if $Pr(S \geq s_k) = P_k$, then the values returned are `qnorm((P_k + p_k)/2)`.

The function `areaProbs` inverts these calculations. If `condmean` is $E[Y|x]$ and `condstd` is $\sqrt{var(Y|x)}$, then this function calculates $Pr(S|x)$ by calculating the area under the normal curve.

Value

For pvecToCutpoints, a vector of length one greater than pvec giving the endpoints of the bins. Note that the first and last values are always infinite.

For pvecToCutpoints, a vector of length the same length as pvec giving the midpoint of the bins.

For areaProbs a vector of probabilities of the same length as pvec.

Warning

Variables are given from *lowest* to *highest* state, for example ‘Low’, ‘Medium’, ‘High’. StatShop expects variables in the opposite order.

Note

The function effectiveThetas does something similar, but assumes all probability values are equally weighted.

Author(s)

Russell Almond

References

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (Draft) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.

Almond, R.G. (Draft) Building Bayesian Network Proficiency Models from Factor Analysis Results. ETS Research Memorandum (not yet submitted).

See Also

[effectiveThetas](#)

Examples

```
probs <- c(Low=.05,Med=.9,High=.05)
cuts <- pvecToCutpoints(probs)
mids <- pvecToMidpoints(probs)
```

```
areaProbs(probs,1,.5)
```

betaci

*Credibility intervals for a proportion based on beta distribution***Description**

This generates upper and lower bounds for a highest posterior density credibility interval for a beta distribution by looking at appropriate quantiles of the beta distribution. This is designed to work with sums of classification probabilities.

Usage

```
betaci(sumData, totals = NULL, limits = c(lower = 0.025, upper = 0.975),
      a = 0.5, b = 0.5)
```

Arguments

sumData	Counts or averages of proportions. Note these do not need to be integers, sums of classification probabilities work here.
totals	Total number of individuals as reference for sumData. If missing or NULL then the value use is colSums(data).
limits	The upper and lower credibility limits.
a	Value for the shape1 parameter of the beta prior.
b	Value for the shape2 parameter of the beta prior.

Details

This function computes the upper and lower bounds of a credibility interval for a beta distribution based on sumData successes out of totals trials. Note that as a beta distribution is used for the basic calculations, neither sumData nor totals need be integers.

To avoid problems with zero cells (or cells with values equal to totals), a small prior is added to the beta calculations. By default a Jeffrey's prior (.5, .5) is added to the data. Thus the final returned value is:

$$qbeta(prob, sumData + a, totals - sumData + b)$$

where prob varies over the values in limits. Note that a and b can be scalars or an array conformable with totals.

Value

A list of the same length as limits with the same names. Each component is a quantile of the posterior distribution which has the same shape as sumData.

Note that limits is not limited to length 2, although this is the most obvious application.

Author(s)

Russell Almond

See Also

See [OCP](#) for an application.

Examples

```
x <- matrix(c(7,4,2,31),2,2)
## Use column sums as totals
betaci(x)

## fixed totals
nn <- matrix(c(30,15,20,35),2,2)
betaci(x,nn)

## Prior varies according to cell.
pi0 <- c(.2,.2,.2,.8)
betaci(x,nn,a=pi0,b=1-pi0)
```

buildFactorTab	<i>Builds probability tables from Scored Bayes net output.</i>
----------------	--

Description

Looks for margin statistics in scored Bayes net output, and puts them into tables with rows representing variables and columns representing variable states.

The marginTab function does this for a single individual. The buildMarginTab uses the grand mean across all individuals and the buildFactorTab breaks down groups according to a factor variable. The function build2FactorTab builds a three-way table.

Usage

```
buildFactorTab(data, fact, stateNames, skillNames, reverse = TRUE,
               stem="margin", sep=".")
build2FactorTab(data, fact1, fact2, stateNames, skillNames,
                reverse = TRUE, stem="margin",sep=".")
buildMarginTab(data, stateNames, skillNames, reverse = TRUE,
               stem="margin",sep=".")
marginTab(datarow, stateNames, skillNames, reverse = TRUE,
          stem="margin",sep=".")
```

Arguments

data	A data sets of StatShop statistics for many individuals.
datarow	One row of such a data set.
fact	A factor variable according to which to split the data. Length should be the same as the length of data.

<code>fact1</code>	A factor variable according to which to split the data.
<code>fact2</code>	A factor variable according to which to split the data.
<code>stateNames</code>	Names of the variable states.
<code>skillNames</code>	Names of the proficiency variable(s) to be used.
<code>reverse</code>	Reverse the order of the states for display (i.e., convert from StatShop order of highest first to more natural order of lowest first.
<code>stem</code>	A character string giving a prefix used to indicate variable names.
<code>sep</code>	A character string giving a separator used to separate prefix from variable names.

Details

This looks for columns marked “<stem><sep><skillName>” in the data frame, and builds them into a matrix. It is assumed that all variables have the same number of states and that they are in the same order, and the order is the same as given in `stateNames`.

The functions `buildFactorTab` and `build2FactorTab` really expect their `skillNames` argument to be a single variable name. However, they should work with multiple variables if suitable values are chosen for the state names.

Value

For `marginTab` a matrix with columns corresponding to `skillNames` and rows corresponding to `stateNames` giving the probabilities for a single individual.

For `buildMarginTab` a matrix with columns corresponding to `skillNames` and rows corresponding to `stateNames` giving the average probabilities for the entire data set.

For `buildFactorTab` a matrix with columns corresponding to the unique values of `fact` and rows corresponding to `stateNames` entries give the average probabilities across the groups.

For `build2FactorTab` a 3 dimensional array with the first dimension corresponding to the unique values of `fact1`, the second dimension corresponding to the unique values of `fact2` and the last dimension corresponding to `stateNames` entries give the average probabilities across the groups.

Author(s)

Russell Almond

See Also

[stackedBars](#), [compareBars](#)

Examples

```
data(ACED)

marginTab(ACED.scores[1,], c("H", "M", "L"), ACED.skillNames, reverse = TRUE,
          stem="p", sep="")

buildMarginTab(ACED.scores, c("H", "M", "L"), ACED.skillNames[1:4],
```



```

reverse = TRUE,
stem="p",sep="")

buildFactorTab(ACED.scores, ACED.scores$Cond, c("H","M","L"), "sgp",
reverse = TRUE,
stem="p", sep="")

build2FactorTab(ACED.scores, ACED.scores$Seq, ACED.scores$FB,
c("H","M","L"), "sgp",
reverse = TRUE, stem="p",sep="")

```

buildParentList	<i>Builds a list of parents of nodes in a graph</i>
-----------------	---

Description

Takes an incidence matrix describing a graph, and an order of the nodes, and builds a list of parents for each node. If the `ord` argument is not supplied, the ordering is constructed through maximum cardinality search (see [mcSearch](#)).

Usage

```
buildParentList(sm, start = colnames(sm)[1], ord = mcSearch(sm, start))
```

Arguments

<code>sm</code>	A logical matrix whose rows and columns correspond to nodes (variables) and a true value indicates an edge between the variables.
<code>start</code>	The name of the first element.
<code>ord</code>	A vector of size equal to the number of columns of <code>sm</code> giving the ordering of nodes.

Details

The `sm` argument should be an incidence matrix for a graph, with row and column names set to the names of the nodes/variables.

A node i is a *parent* of node j if

- a They are neighbors, that is `sm[i, j] == TRUE`.
- b The node i appears before node j in `ord`, that is `ord[i] < ord[j]`.

The argument `start` is used only if `ord` is not supplied, in which case it is passed to `mcSearch`.

Value

A list with as many elements as there are columns in `sm`, and whose elements appear in the order specified by `ord`. Each element of that list is a character vector giving the names of the parents.

Author(s)

Russell Almond

References

Almond, R. G. (2010). 'I can name that Bayesian network in two matrixes.' *International Journal of Approximate Reasoning*. **51**, 167-178.

See Also

[mcSearch](#), [structMatrix](#)

Examples

```
data(MathGrades)
MG.struct <- structMatrix(MathGrades$var)

parents <- buildParentList(MG.struct) # Arbitrary start
parentsa <- buildParentList(MG.struct, "Algebra") # Put algebra first.
```

buildRegressions	<i>Creates a series of regressions from a covariance matrix</i>
------------------	---

Description

This function takes a covariance matrix and list of variables and their parents and returns a collection of regression model parameters for each variable regressed on its parents. This is a compact representation of a normal graphical model.

Usage

```
buildRegressions(Sigma, means = 0,
                 parents = buildParentList(structMatrix(Sigma)))
```

Arguments

Sigma	A covariance matrix among a collection of continuous variables.
means	The means of those variables
parents	A named list of length equal to <code>ncol(Sigma)</code> whose elements should correspond to the rows/columns of Sigma. Each element should be a character vector giving the names of the parents of the given node.

Details

This function performs one regression for each element of the parents list. The name of the dependent variable for each regression is given by `names(parents)` and the independent variables is given by the values of parents. (The function `buildParentList()` builds a suitable list of elements.)

If means is not supplied, then the variables are assumed to be centered, otherwise the given vector is used as the means.

Value

A list of length equal to parents whose elements are also a list having the following structure

<code>b</code>	A vector of slopes for the regression with names equal to the names of the parent variables. Note that if there are no parents, this will be <code>numeric(0)</code> .
<code>a</code>	The intercept from the regression.
<code>std</code>	The residual standard deviation from the regression.

Author(s)

Russell Almond

References

Almond, R. G. (2010). 'I can name that Bayesian network in two matrixes.' *International Journal of Approximate Reasoning*. **51**, 167-178.

See Also

[buildParentList](#), [buildRegressionTables](#)

Examples

```
data(MathGrades)
pl <- buildParentList(structMatrix(MathGrades$var),"Algebra")
rt <- buildRegressions(MathGrades$var,MathGrades$mean,pl)
```

`buildRegressionTables` *Builds conditional probability tables from regressions*

Description

Takes a description of a normal graphical model as a series of regressions and a description of the corresponding discrete variables and builds a series of conditional probability tables for the corresponding Bayesian network.

Usage

```
buildRegressionTables(regs, pvecs, mean = 0, std = 1)
```

Arguments

<code>regs</code>	A list with names corresponding to the variables in the model giving a series of regression coefficients (see Details).
<code>pvecs</code>	A list with names corresponding to the variables in the model giving a series of probability vectors in order from highest state to lowest state (see Details).
<code>mean</code>	A vector of means for the continuous variables.
<code>std</code>	A vector of standard deviations for the continuous variables.

Details

The `regs` argument should be a list whose names are the names of the variables. Each element should have the following fields:

- b** A vector of slopes for the regression with names equal to the names of the parent variables. Note that if there are no parents, this should be `numeric(0)`.
- a** The intercept from the regression.
- std** The residual standard deviation from the regression.

The function `buildRegressions()` creates an appropriate list.

The `pvecs` should be a list whose names are the names of the variables. Each element should be a named vector of probabilities in order from the *Highest* to the *Lowest* state, e.g. `c(High=.2, Med=.5, Low=.3)`.

The values `mean` and `std` should either be scalars or vectors of the same length as the number of elements in `regs` and `pvecs`. If vectors, they should have names corresponding to the variable names. Note that the order of the elements does not need to be the same in all four arguments, but that the names of all four arguments need to be identical (unless `mean` or `std` are given as scalars, in which case they will be appropriately replicated.)

Value

A list of conditional probability tables whose names correspond to `regs`. Each conditional probability table is expressed as a data frame whose columns correspond to either parent variables or states of the child variable and whose rows correspond to configurations of the parent variable.

Warning

Variables are given from *highest* to *lowest* state, for example 'High', 'Medium', 'Low'. This is the order expected by StatShop. Note that `pvecToCutpoints` expects probability vectors in the opposite order.

Author(s)

Russell Almond

References

Almond, R. G. (2010). ‘I can name that Bayesian network in two matrixes.’ *International Journal of Approximate Reasoning*. **51**, 167-178.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

See Also

[buildRegressions](#)

Examples

```
data(MathGrades)
pl <- buildParentList(structMatrix(MathGrades$var), "Algebra")
rt <- buildRegressions(MathGrades$var, MathGrades$means, pl)
tabs <- buildRegressionTables(rt, MathGrades$pvecs, MathGrades$means,
                             sqrt(diag(MathGrades$var)))
```

calcDDTable	<i>Calculates DiBello–Dirichlet model probability and parameter tables</i>
-------------	--

Description

The DiBello–Dirichlet model creates a hyper-Dirichlet prior distribution by interpolating between an masterProfile and a noviceProfile. This function builds the hyper-Dirichlet parameter table, or with normalization, the conditional probability table for this distribution type.

Usage

```
calcDDTable(skillLevels, obsLevels, skillWeights, masterProfile,
            noviceProfile = 0.5, rule = "Compensatory")
calcDDFrame(skillLevels, obsLevels, skillWeights, masterProfile,
            noviceProfile = 0.5, rule = "Compensatory")
```

Arguments

- | | |
|---------------|--|
| skillLevels | A list of character vectors giving names of levels for each of the condition variables. |
| obsLevels | A character vector giving names of levels for the output variables from highest to lowest. As a special case, can also be a vector of integers. |
| skillWeights | A numeric vector of the same length as skillLevels giving the weight to be applied to each skill. |
| masterProfile | The Dirichlet prior for “experts” (see Details). Its length should match obsLevels. |
| noviceProfile | The Dirichlet prior for “novices” (see Details). Its length should match obsLevels or as a special case a scalar quantity gives a uniform prior. Default is uniform prior with weight 1/2. |
| rule | Function for computing effective theta (see Details). |

Details

Assume for the moment that there are two possible skill profiles: “expert” and “novice”. This model presumes a conditional probability table for the outcome given skill profile with two rows each of which is an independent categorical distribution. The natural conjugate prior is an independent Dirichlet distribution for each row. The parameters for this distribution are given in the `masterProfile` and `noviceProfile` arguments.

If there is more than one parent variable or if the parent variable has more than one state, the situation becomes muddier. The “expert” state is obviously the one with all the variables at the highest levels and the “novice” is the one with all variables at the lowest levels. If we can assign an integer between 0 and 1 to each of the intermediate states, then we can interpolate between them to produce Dirichlet priors for each row.

This distribution type uses the DiBello effective theta technique to come up with the interpolation. Each parent variable state is assigned a ‘theta’ value using the `effectiveThetas` function to assign a numeric value to each one. These are then combined using the function `rule` in the `rule` argument. The resulting theta values are then scaled to a range of 0–1. The prior for that row is a weighted combination of the `masterProfile` and `noviceProfile`.

The combination of the individual effective theta values into a joint value for effective theta is done by the function `reference by rule`. This should be a function of three arguments: `theta` — the vector of effective theta values for each parent, `alphas` — the vector of discrimination parameters, and `beta` — a scalar value giving the difficulty. The initial distribution supplies three functions appropriate for use with `calcDSTable`: `Compensatory`, `Conjunctive`, and `Disjunctive`. Note that the `beta` argument is effectively ignored because of the later scaling of the output.

Normally `obslevel` should be a character vector giving state names. However, in the special case of state names which are integer values, R will “helpfully” convert these to legal variable names by prepending a letter. This causes other functions which rely on the `names()` of the result being the state names to break. As a special case, if the value of `obsLevel` is of type numeric, then `calcDSFrame()` will make sure that the correct values are preserved.

Value

For `calcDDTable`, a matrix whose rows correspond configurations of the parent variable states (`skillLevels`) and whose columns correspond to `obsLevels`. Each row of the table is the parameters of a Dirichlet distribution, so the whole matrix is the parameters for a hyper-Dirichlet distribution. The order of the parent rows is the same as is produced by applying `expand.grid` to `skillLevels`.

For `calcDDFrame` a data frame with additional columns corresponding to the entries in `skillLevels` giving the parent value for each row.

Note

Unlike `calcDSTable`, there is not a corresponding DiBello-Dirichlet distribution support in StatShop. This function is used to model the parameters of an unconstrained hyper-Dirichlet distribution.

This was originally designed for use in Situational Judgment Tests where experts might not agree on the “key”.

Note: Zeros in the `masterProfile` indicate responses that a master would **never** make. They will result in zero probability of mastery for any response which yields that outcome.

References

Almond, R.G. and Roberts, R. (Draft) Bayesian Scoring for Situational Judgment Tests. Unpublished white paper.

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (Draft) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.

Almond, R.G., DiBello, L., Jenkins, F., Mislevy, R.J., Senturk, D., Steinberg, L.S. and Yan, D. (2001) Models for Conditional Probability Tables in Educational Assessment. *Artificial Intelligence and Statistics 2001* Jaakkola and Richardson (eds)., Morgan Kaufmann, 137–143.

See Also

[effectiveThetas](#), [Compensatory](#), [calcDNTable](#), [calcDNSTable](#), [expand.grid](#)

Examples

```
skill11 <- c("High", "Medium", "Low")
skill21 <- c("High", "Low")
option5L <- c("A", "B", "C", "D", "E")

## Expert responses
eProfile <- c(A=7, B=15, C=3, D=0, E=0)

paramT <- calcDDTable(list(S1=skill11, S2=skill21), option5L,
                        c(S1=2, S2=1), masterProfile=eProfile+0.5)

paramF <- calcDNFrame(list(S1=skill11, S2=skill21), option5L,
                        c(S1=2, S2=1), masterProfile=5*eProfile+0.5,
                        noviceProfile=2)
```

calcDNTable

Creates the probability table for DiBello–Normal distribution

Description

The calcDNTable function takes a description of input and output variables for a Bayesian network distribution and a collection of IRT-like parameter (discrimination, difficulty) and calculates a conditional probability table using the DiBello–Normal distribution (see Details). The calcDNFrame function returns the value as a data frame with labels for the parent states.

Usage

```
calcDNTable(skillLevels, obsLevels, lnAlphas, beta, std, rule = "Compensatory")
calcDNFrame(skillLevels, obsLevels, lnAlphas, beta, std, rule = "Compensatory")
```

Arguments

skillLevels	A list of character vectors giving names of levels for each of the condition variables.
obsLevels	A character vector giving names of levels for the output variables from highest to lowest. Can also be a vector of integers (see Details).
lnAlphas	A vector of log slope parameters. Its length should be either 1 or the length of skillLevels, depending on the choice of rule.
beta	A vector of difficulty (-intercept) parameters. Its length should be either 1 or the length of skillLevels, depending on the choice of rule.
std	The log of the residual standard deviation (see Details).
rule	Function for computing effective theta (see Details).

Details

The DiBello–Normal distribution (Almond et al, Draft) is a variant of the DiBello–Samejima distribution (Almond et al, 2001) for creating conditional probability tables for Bayesian networks which uses a regression-like (probit) link function in place of Samejima’s graded response link function. The basic procedure unfolds in three steps.

1. Each level of each input variable is assigned an “effective theta” value — a normal value to be used in calculations.
2. For each possible skill profile (combination of states of the parent variables) the effective thetas are combined using a combination function. This produces an “effective theta” for that skill profile.
3. Taking the effective theta value as the mean, the probability that the examinee will fall into each category.

The parent (conditioning) variables are described by the skillLevels argument which should provide for each parent variable in order the names of the states ranked from highest to lowest value. These are calculated through the function `effectiveThetas` which gives equally spaced points on the probability curve. Note that for the DiBello–Normal distribution, Step 1 and Step 3 are inverses of each other (except for rounding error).

The combination of the individual effective theta values into a joint value for effective theta is done by the function reference by rule. This should be a function of three arguments: theta — the vector of effective theta values for each parent, alphas — the vector of discrimination parameters, and beta — a scalar value giving the difficulty. The initial distribution supplies five functions appropriate for use with calcDNTable: `Compensatory`, `Conjunctive`, and `Disjunctive`, `OffsetConjunctive`, and `OffsetDisjunctive`. The last two have a slightly different parameterization: alpha is assumed to be a scalar and betas parameter is vector valued. Note that the discrimination and difficulty parameters are built into the structure function and not the probit curve.

The effective theta values are converted to probabilities by assuming that the categories for the consequence variable (obsLevels) are generated by taking equal probability divisions of a standard normal random variable. However, a person with a given pattern of condition variables is drawn from a population with mean at effective theta and standard deviation of $\exp(\text{std})$. The returned numbers are the probabilities of being in each category.

Normally `obslevel` should be a character vector giving state names. However, in the special case of state names which are integer values, R will “helpfully” convert these to legal variable names by prepending a letter. This causes other functions which rely on the `names()` of the result being the state names to break. As a special case, if the value of `obsLevel` is of type `numeric`, then `calcDNFrame()` will make sure that the correct values are preserved.

Value

For `calcDNTTable`, a matrix whose rows correspond configurations of the parent variable states (`skillLevels`) and whose columns correspond to `obsLevels`. Each row of the table is a probability distribution, so the whole matrix is a conditional probability table. The order of the parent rows is the same as is produced by applying `expand.grid` to `skillLevels`.

For `calcDNFrame` a data frame with additional columns corresponding to the entries in `skillLevels` giving the parent value for each row.

Note

This distribution class was developed primarily for modeling relationships among proficiency variables. For models for observables, see [calcDSTable](#).

Author(s)

Russell Almond

References

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (Draft) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.

Almond, R.G., DiBello, L., Jenkins, F., Mislevy, R.J., Senturk, D., Steinberg, L.S. and Yan, D. (2001) Models for Conditional Probability Tables in Educational Assessment. *Artificial Intelligence and Statistics 2001* Jaakkola and Richardson (eds).. Morgan Kaufmann, 137–143.

See Also

```
effectiveThetas,Compensatory,OffsetConjunctive,eThetaFrame,calcDSTable,calcDNllike,
calcDPCTable.expand.grid
```

Examples

[illegible]

```
rule="Compensatory")
```

calcDPCTable

Creates the probability table for the discrete partial credit model

Description

The calcDPCTable function takes a description of input and output variables for a Bayesian network distribution and a collection of IRT-like parameter (discrimination, difficulty) and calculates a conditional probability table using the discrete partial credit distribution (see Details). The calcDPCFrame function returns the value as a data frame with labels for the parent states.

Usage

```
calcDPCTable(skillLevels, obsLevels, lnAlphas, betas,
             rules = "Compensatory", link="partialCredit",
             linkScale=NULL, Q=TRUE,
             tvals=lapply(skillLevels,
                          function (sl) effectiveThetas(length(sl))))

calcDPCFrame(skillLevels, obsLevels, lnAlphas, betas,
             rules = "Compensatory", link="partialCredit",
             linkScale=NULL, Q=TRUE,
             tvals=lapply(skillLevels,
                          function (sl) effectiveThetas(length(sl))))
```

Arguments

- | | |
|-------------|---|
| skillLevels | A list of character vectors giving names of levels for each of the condition variables. |
| obsLevels | A character vector giving names of levels for the output variables from highest to lowest. As a special case, can also be a vector of integers. |
| lnAlphas | A list of vectors of log slope parameters. Its length should be 1 or length(obsLevels)-1. The required length of the individual component vectors depends on the choice of rule (and is usually either 1 or the length of skillLevels). |
| betas | A list of vectors of difficulty (-intercept) parameters. Its length should be 1 or length(obsLevels)-1. The required length of the individual component vectors depends on the choice of rule (and is usually either 1 or the length of skillLevels). |
| rules | A list of functions for computing effective theta (see Details). Its length should be length(obsLevels)-1 or 1 (implying that the same rule is applied for every gap.) |
| link | The function that converts a table of effective thetas to probabilities |
| linkScale | An optional scale parameter for the link function. This is only used with certain choices of link function. |

Q	This should be a Q matrix indicating which parent variables are relevant for which state transitions. It should be a number of states minus one by number of parents logical matrix. As a special case, if all variable are used for all levels, then it can be a scalar value.
tvals	A list of the same length as skillLevels. Each element should be a numeric vector values on the theta (logistic) scale corresponding to the levels for that parent variable. The default spaces them equally according to the normal distribution (see effectiveThetas).

Details

The discrete graded response model is a generalization of the DiBello–Samejima mechanism for creating conditional probability tables for Bayesian network models using IRT-like parameters ([calcDSTable](#)). The basic procedure unfolds in three steps.

1. Each level of each input variable is assigned an “effective theta” value — a normal value to be used in calculations.
2. For each possible skill profile (combination of states of the parent variables) the effective thetas are combined using a one of the rule functions. This produces an “effective theta” for that skill profile.
3. The effective theta table is input into the link function to produce a probability distribution over the states of the outcome variables.

The parent (conditioning) variables are described by the `skillLevels` argument which should provide for each parent variable in order the names of the states ranked from highest to lowest value. The default implementation uses the function [effectiveThetas](#) to calculate equally spaced points on the normal curve. This can be overridden by supplying a `tvals` argument. This should be a list of the same length as `skillLevels` with each element having the same length as the corresponding element of `skillLevels`.

The `tvals` (either default or user supplied) are used to create a table of rows with values $\theta_1, \dots, \theta_K$, corresponding to all possible combinations of the parent variables (using [expand.grid](#)).

Let X be the child variable of the distribution, and assume that it can take on M possible states labeled x_1 through x_M in increasing order. (Note: that `calcDPCTable` assumes variable states are ordered the other direction: from highest to lowest.) For each state but the lowest state (the last one in the input order) defines a combination rule $Z_m(\theta_1, \dots, \theta_K; \text{alphas}, \text{betas})$. Applying these functions to the rows of the table produces a table of effective thetas for each configuration of the parent variables and each child state except for the lowest. (The metaphor is this theta represents the “ability level” required to reach that output state.)

Note that the $Z_m(\cdot)$ s do not need to have the same parameters or even the same functional form. The argument `rules` should contain a list of the names of the combination functions, the first one corresponding to $Z_M(\cdot)$, and so forth in descending order. As a special case, if `rules` has only one element, then it is used for all of the transitions. Similarly, the `lnAlphas` and `betas` should also be lists of the parameters of the combination functions corresponding to the transitions between the levels. The `betas[[m]]` represent difficulties (negative intercepts) and the `exp(lnAlphas[[m]])` represent slopes for the transition to level m (following the highest to lowest order). Again if these lists have length one, the same value is used for all transitions.

The length of the elements of `lnAlphas` and `betas` is determined by the specific choice of combination function. The functions [Compensatory](#), [Conjunctive](#), and [Disjunctive](#) all assume that there

will be one `lnAlpha` for each parent variable, but a single `beta`. The functions `OffsetConjunctive`, and `OffsetDisjunctive` both assume that there will be one `beta` for each parent variable, but a single `lnAlpha`.

The code `link` function is then applied to the table of effective `theta` values to produce a conditional probability distribution. Two link functions are currently supported: `partialCredit` is based on the generalized partial credit model (Muraki, 1992), `gradedResponse` is a modified version of the graded response model (Samejima, 1969). (The modification corrects for problems when the curves cross.) A third planned link function is based on a normal error model, this will require the extra `linkScale` parameter.

The `Q` matrix is used in situations where some of the parent variables are not relevant for one or more parent transitions. If parent `k` is relevant for the transition between state `m+1` and `m` (remember that states are coded from highest to lowest) then `Q[m, k]` should be `TRUE`. In particular, `eTheta[, Q[m,]]` is passed to the combination rule, not all of `theta`. If there are false entries in `Q` the corresponding sets of `alphas` and `betas` need to have the correct length. Generally speaking, `Q` matrixes with `FALSE` entries are not appropriate with the `gradedResponse` link. As a special case if `Q=TRUE`, then all parent variables are used for all state transitions.

Normally `obslevel` should be a character vector giving state names. However, in the special case of state names which are integer values, R will “helpfully” convert these to legal variable names by prepending a letter. This causes other functions which rely on the `names()` of the result being the state names to break. As a special case, if the value of `obslevel` is of type `numeric`, then `calcDSFrame()` will make sure that the correct values are preserved.

Value

For `calcDPCTable`, a matrix whose rows correspond configurations of the parent variable states (`skillLevels`) and whose columns correspond to `obsLevels`. Each row of the table is a probability distribution, so the whole matrix is a conditional probability table. The order of the parent rows is the same as is produced by applying `expand.grid` to `skillLevels`.

For `calcDPCFrame` a data frame with additional columns corresponding to the entries in `skillLevels` giving the parent value for each row.

Note

The framework set up by this function is completely expandable. The `link` and the elements of rules can be any value that is suitable for the first argument of `do.call`.

Elements of rules are called with the expression `do.call(rules[[kk]],list(thetas,exp(lnAlphas[[kk]]),betas[[kk]]))` where `thetas` is the matrix of effective `theta` values produced in the first step of the algorithm, and the return function should be a vector of effective `thetas`, one for each row of `thetas`.

The link function is called with the expression `do.call(link,list(et,linkScale,obsLevels))` where `et` is the matrix of effective `thetas` produced in the second step. It should return a conditional probability table with the same number of rows and one more column than `et`. All of the rows should sum to 1.0.

Author(s)

Russell Almond

References

- Almond, R.G. (2015). An IRT-based Parameterization for Conditional Probability Tables. Paper submitted to the 2015 Bayesian Application Workshop at the Uncertainty in Artificial Intelligence conference.
- Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (2015) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.
- Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, **16**, 159-176. DOI: 10.1177/014662169201600206
- Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, **34**, (No. 4, Part 2).

See Also

[effectiveThetas](#), [Compensatory](#), [OffsetConjunctive](#), [eThetaFrame](#), [calcDNTable](#), [calcDSTable](#), [expand.grid](#), [gradedResponse](#), [partialCredit](#)

Examples

```
## Set up variables
skill1l <- c("High", "Medium", "Low")
skill2l <- c("High", "Medium", "Low", "LowerYet")
correctL <- c("Correct", "Incorrect")
pcreditL <- c("Full", "Partial", "None")
gradeL <- c("A", "B", "C", "D", "E")

## Simple binary model, these three should be the same.
cptCorrect <- calcDPCTable(list(S1=skill1l, S2=skill2l), correctL,
                             log(c(S1=1, S2=.75)), 1.0, rule="Compensatory",
                             link="partialCredit")
cptCorrect2 <- calcDPCTable(list(S1=skill1l, S2=skill2l), correctL,
                             log(c(S1=1, S2=.75)), 1.0, rule="Compensatory",
                             link="gradedResponse")
cptCorrect1 <- calcDSTable(list(S1=skill1l, S2=skill2l), correctL,
                           log(c(S1=1, S2=.75)), 1.0, rule="Compensatory")
stopifnot (all (abs(cptCorrect2-cptCorrect1) <.001))
stopifnot (all (abs(cptCorrect-cptCorrect1) <.001))

## Conjunctive uses multiple betas, not multiple alphas.
cptConj <- calcDPCTable(list(S1=skill1l, S2=skill2l), correctL,
                        log(1), c(S1=0.5, S2=.7), rule="OffsetConjunctive")

## Simple model, Skill 1 needed for step 1, Skill 2 for Step 2.
cptPC1 <- calcDPCFrame(list(S1=skill1l, S2=skill2l), pcreditL,
                      lnAlphas=log(1),
                      betas=list(full=c(S1=0, S2=999), partial=c(S2=999, S2=0)),
                      rule="OffsetDisjunctive")

## Variant using Q-matrix
cptPC1a <- calcDPCTable(list(S1=skill1l, S2=skill2l), pcreditL,
                        lnAlphas=log(1),
```

```

      betas=list(full=c(S1=0),partial=c(S2=0)),
      Q=matrix(c(TRUE,FALSE,FALSE,TRUE),2,2),
      rule="OffsetDisjunctive")
stopifnot(all(abs(as.vector(numericPart(cptPC1))-as.vector(cptPC1a))<.0001))

## Complex model, different rules for different levels
cptPC2 <- calcDPCTable(list(S1=skill1l1,S2=skill1l2),pcreditL,
      list(full=log(1),partial=log(c(S1=1,S2=.75))),
      betas=list(full=c(0,999),partial=1.0),
      rule=list("OffsetDisjunctive","Compensatory"))

## Graded Response Model, typically uses different difficulties
cptGraded <- calcDPCTable(list(S1=skill1l1),gradeL,
      log(1),betas=list(A=2,B=1,C=0,D=-1),
      rule="Compensatory",link="gradedResponse")

## Partial credit link is somewhat different
cptPC5 <- calcDPCTable(list(S1=skill1l1),gradeL,
      log(1),betas=list(A=2,B=1,C=0,D=-1),
      rule="Compensatory",link="partialCredit")
cptPC5a <- calcDPCTable(list(S1=skill1l1),gradeL,
      log(1),betas=1,
      rule="Compensatory",link="partialCredit")

## Need to be careful when using different slopes (or non-increasing
## difficulties) with graded response link as curves may cross.

cptCross <- calcDPCTable(list(S1=skill1l1),pcreditL,
      log(1),betas=list(full=-1,partial=1),
      rule="Compensatory",link="gradedResponse")
stopifnot (all(abs(cptCross[, "Partial"])<.001))

```

calcDSlike	<i>Calculates the log-likelihood for data from a DiBello–Samejima (Normal) distribution</i>
------------	---

Description

These functions take data which represent draws from a categorical data with the given DiBello–Samejima distribution and returns the log-likelihood of the given data.

Usage

```

calcDSlike(data, parents, skillLevels, child, obsLevels,
  lnAlphas, beta, dinc = 0, rule = "Compensatory")
calcDNlike(data, parents, skillLevels, child, obsLevels,
  lnAlphas, beta, std, rule = "Compensatory")

```

Arguments

data	A data frame whose columns contain variables corresponding to parent and child.
parents	A vector of names for the columns in data corresponding to the parent variables.
child	The name of the child variable, should refer to a column in data.
skillLevels	A list of character vectors giving names of levels for each of the condition variables.
obsLevels	A character vector giving names of levels for the output variables from highest to lowest.
lnAlphas	A vector of log slope parameters. Its length should be either 1 or the length of skillLevels, depending on the choice of rule.
beta	A vector of difficulty (-intercept) parameters. Its length should be either 1 or the length of skillLevels, depending on the choice of rule.
dinc	Vector of difficulty increment parameters (see calcDSTable).
rule	Function for computing effective theta (see calcDSTable).
std	The log of the residual standard deviation (see Details).

Details

This function assumes that the observed data are independent draws from a Bayesian network. This function calculates the log-likelihood of a single conditional probability table. First, it calculates a table of counts corresponding states of the parent and child variables using the function [dataTable](#). Next it calculates the conditional probability for each cell using the function [calcDSTable](#) or [calcDNTable](#).

It then calculates the log-likelihood as the sum of $count(cell) * \log(Pr(cell))$ where this value is set to zero if $count(cell)$ is zero (this allows cells with zero probability as long as the count is also zero).

Value

A real giving the log-likelihood of the observed data.

Note

This function is primarily about testing the log likelihood calculations used internally in StatShop.

Author(s)

Russell Almond

References

<http://comet.research.ets.org/~ralmond/StatShop>

See Also

[dataTable](#), [calcDSTable](#), [CompensatoryOffsetConjunctive](#), [eThetaFrame](#), [calcDNTable](#)

Examples

```
skill1l1 <- c("High", "Medium", "Low")
skill3l1 <- c("High", "Better", "Medium", "Worse", "Low")
correctL <- c("Correct", "Incorrect")

x <- read.csv(paste(library(help="CPTtools")$path, "testFiles",
  "randomPinned100.csv", sep=Platform$file.sep),
  header=FALSE, as.is=TRUE,
  col.names = c("Skill1", "Skill2", "Skill3",
    "Comp.Correct", "Comp.Grade",
    "Conj.Correct", "Conj.Grade",
    "Cor.Correct", "Cor.Grade",
    "Dis.Correct", "Dis.Grade",
    "Inhib.Correct", "Inhib.Grade"
  ))
x[, "Skill1"] <- ordered(x[, "Skill1"], skill1l1)
x[, "Skill3"] <- ordered(x[, "Skill3"], skill3l1)
x[, "Comp.Correct"] <- ordered(x[, "Comp.Correct"], correctL)

like <- calcDSlike(x, c("Skill1", "Skill3"),
  list(Skill1=skill1l1, Skill3=skill3l1),
  "Comp.Correct", correctL,
  log(c(0.45, -0.4)), -1.9, rule="Compensatory")
```

calcDSTable

Creates the probability table for DiBello–Samejima distribution

Description

The `calcDSTable` function takes a description of input and output variables for a Bayesian network distribution and a collection of IRT-like parameter (discrimination, difficulty) and calculates a conditional probability table using the DiBello–Samejima distribution (see Details). The `calcDSFrame` function returns the value as a data frame with labels for the parent states.

Usage

```
calcDSTable(skillLevels, obsLevels, lnAlphas, beta, dinc = 0,
  rule = "Compensatory")
calcDSFrame(skillLevels, obsLevels, lnAlphas, beta, dinc = 0,
  rule = "Compensatory")
```

Arguments

skillLevels A list of character vectors giving names of levels for each of the condition variables.

obsLevels	A character vector giving names of levels for the output variables from highest to lowest. As a special case, can also be a vector of integers.
lnAlphas	A vector of log slope parameters. Its length should be either 1 or the length of skillLevels, depending on the choice of rule.
beta	A vector of difficulty (-intercept) parameters. Its length should be either 1 or the length of skillLevels, depending on the choice of rule.
dinc	Vector of difficulty increment parameters (see Details).
rule	Function for computing effective theta (see Details).

Details

The DiBello–Samejima model is a mechanism for creating conditional probability tables for Bayesian network models using IRT-like parameters. The basic procedure unfolds in three steps.

1. Each level of each input variable is assigned an “effective theta” value — a normal value to be used in calculations.
2. For each possible skill profile (combination of states of the parent variables) the effective thetas are combined using a combination function. This produces an “effective theta” for that skill profile.
3. The effective theta is input into Samejima’s graded-response model to produce a probability distribution over the states of the outcome variables.

The parent (conditioning) variables are described by the skillLevels argument which should provide for each parent variable in order the names of the states ranked from highest to lowest value. The original method (Almond et al., 2001) used equally spaced points on the interval $[-1, 1]$ for the effective thetas of the parent variables. The current implementation uses the function [effectiveThetas](#) to calculate equally spaced points on the normal curve.

The combination of the individual effective theta values into a joint value for effective theta is done by the function reference by rule. This should be a function of three arguments: theta — the vector of effective theta values for each parent, alphas — the vector of discrimination parameters, and beta — a scalar value giving the difficulty. The initial distribution supplies five functions appropriate for use with calcDSTable: [Compensatory](#), [Conjunctive](#), and [Disjunctive](#), [OffsetConjunctive](#), and [OffsetDisjunctive](#). The last two have a slightly different parameterization: alpha is assumed to be a scalar and betas parameter is vector valued. Note that the discrimination and difficulty parameters are built into the structure function and not the IRT curve.

The Samejima graded response link function describes a series of curves:

$$P_m(\theta) = Pr(X \geq x - m|\theta) = \text{logit}^{-1}(\theta - d_m)$$

for $m > 1$, where $D = 1.7$ (a scale factor to make the logistic curve match more closely with the probit curve). The probability for any given category is then the difference between two adjacent logistic curves. Note that because a difficulty parameter was included in the structure function, we have the further constraint that $\sum d_m = 0$.

To remove the parameter restriction we work with the difference between the parameters: $d_m - d_{m-1}$. The value of d_2 is set at $-\text{sum(dinc)}/2$ to center the d values. Thus the dinc parameter (which is required only if $\text{length(obsLevels)} > 2$) should be of length $\text{length(obsLevels)} - 2$. The first value is the difference between the d values for the two highest states, and so forth.

Normally `obslevel` should be a character vector giving state names. However, in the special case of state names which are integer values, R will “helpfully” convert these to legal variable names by prepending a letter. This causes other functions which rely on the `names()` of the result being the state names to break. As a special case, if the value of `obsLevel` is of type numeric, then `calcDSFrame()` will make sure that the correct values are preserved.

Value

For `calcDSTable`, a matrix whose rows correspond configurations of the parent variable states (`skillLevels`) and whose columns correspond to `obsLevels`. Each row of the table is a probability distribution, so the whole matrix is a conditional probability table. The order of the parent rows is the same as is produced by applying `expand.grid` to `skillLevels`.

For `calcDSFrame` a data frame with additional columns corresponding to the entries in `skillLevels` giving the parent value for each row.

Note

This distribution class is not suitable for modeling relationship among proficiency variable, primarily because the normal mapping used in the effective theta calculation and the Samejima graded response models are not inverses. For those model, the function `calcDNTable`, which uses a probit link function, is recommended instead.

Author(s)

Russell Almond

References

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (2015) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.

Almond, R.G., DiBello, L., Jenkins, F., Mislevy, R.J., Senturk, D., Steinberg, L.S. and Yan, D. (2001) Models for Conditional Probability Tables in Educational Assessment. *Artificial Intelligence and Statistics 2001* Jaakkola and Richardson (eds), Morgan Kaufmann, 137–143.

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, **34**, (No. 4, Part 2).

See Also

[effectiveThetas](#), [Compensatory](#), [OffsetConjunctive](#), [eThetaFrame](#), [calcDNTable](#), [calcDSlike](#), [calcDPCTable](#), [expand.grid](#)

Examples

```
## Set up variables
skill1l <- c("High", "Medium", "Low")
skill2l <- c("High", "Medium", "Low", "LowerYet")
correctL <- c("Correct", "Incorrect")
gradeL <- c("A", "B", "C", "D", "E")
```

```

cptCorrect <- calcDSTable(list(S1=skill111,S2=skill121),correctL,
                             log(c(S1=1,S2=.75)),1.0,rule="Conjunctive")

cpfCorrect <- calcDSFrame(list(S1=skill111,S2=skill121),correctL,
                           log(c(S1=1,S2=.75)),1.0,rule="Conjunctive")

cptGraded <- calcDSTable(list(S1=skill111),gradeL, 0.0, 0.0, dinc=c(.3,.4,.3))

```

calcNoisyAndTable	<i>Calculate the conditional probability table for a Noisy-And or Noisy-Min distribution</i>
-------------------	--

Description

Calculates the conditional probability table for a noisy-and distribution. This follows a logical model where all inputs must be true for the output to be true; however, some "noise" is allowed that produces random deviations from the pure logic. The noisy-min is a generalization in which all variables are ordered, and the weakest of the parent variables drives the conditional probabilities of the child variable.

Usage

```

calcNoisyAndTable(skillLevels, obsLevels = c("True", "False"),
                  bypass = rep(0, length(skillLevels)), noSlip=1,
                  thresholds = sapply(skillLevels, function(states) states[1]))
calcNoisyAndFrame(skillLevels, obsLevels = c("True", "False"),
                  bypass = rep(0, length(skillLevels)), noSlip=1,
                  thresholds = sapply(skillLevels, function(states) states[1]))

```

Arguments

skillLevels	A list of character vectors giving names of levels for each of the condition variables.
obsLevels	A character vector giving names of levels for the output variables from highest to lowest. As a special case, can also be a vector of integers. Its length should be 2, and the first value is considered to be logically equivalent to "true".
noSlip	A scalar value between 0 and 1. This represents the probability that the output will be true when all of the inputs are true (e.g., 1 - the probability that an examinee will make a careless error).
bypass	A vector of the same length as skillLevels. For each parent variable, this represents the probability that the process will act as if that input condition is met, even if it is not met.
thresholds	If the input variables have more than two states, values that are equal to or higher than this threshold are considered true. It is assumed that the states of the variables are ordered from highest to lowest.

Details

The noisy-and distribution assumes that both the input and output variables are binary. Basically, the output should be true if all of the inputs are true. Let $S_k = 1$ if the k th input is true, and let r_k be the bypass parameter corresponding to the k th input variable. (If the S_k 's represent a skill, then r_k represents the probability that an examinee who lacks that skill will bypass the need for that skill in solving the problem.) Then the probability of the true state for the output variable will be:

$$\Pr(X = \text{True}|\mathbf{S}) = r_0 \prod_k r_k^{1-S_k},$$

where r_0 (the noSlip parameter) is the probability that the output will be true even when all of the inputs are true.

It is assumed that all variables are ordered from highest to lowest state, so that the first state corresponds to "true" the others to false. If the input variable has more than two states, then it can be reduced to a binary variable by using the threshold argument. Any values which are equal to or higher than the threshold for that variable are assumed to be true. (In this case, higher means closer to the the beginning of the list of possible values.)

The noisy-min is a generalization

Value

For calcNoisyAndTable, a matrix whose rows correspond configurations of the parent variable states (skillLevels) and whose columns correspond to obsLevels. Each row of the table is a probability distribution, so the whole matrix is a conditional probability table. The order of the parent rows is the same as is produced by applying expand.grid to skillLevels.

For calcNoisyAndFrame a data frame with additional columns corresponding to the entries in skillLevels giving the parent value for each row.

Note

This is related to the DINA and NIDA models, but uses a slightly different parameterization. In particular, if the noSlip parameter is omitted, it is a noisy input deterministic and-gate (NIDA), and if the bypass parameters are omitted, it is similar to a deterministic input noisy and-gate (DINA), except it lacks a guessing parameter.

Author(s)

Russell Almond

References

- Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (Draft) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.
- Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.
- Diez, F. J. (1993) Parameter adjustment in Bayes networks. The generalized noisy OR-gate. In Heckerman and Mamdani (eds) *Uncertainty in Artificial Intelligence 93*. Morgan Kaufmann. 99–105.

Srinivas, S. (1993) A generalization of the Noisy-Or model, the generalized noisy OR-gate. In Heckerman and Mamdani (eds) *Uncertainty in Artificial Intelligence 93*. Morgan Kaufmann. 208–215.

See Also

[calcDSTable](#), [calcDNTable](#), [calcDPCTable](#), [expand.grid](#), [calcNoisyOrTable](#)

Examples

```
## Logical and table
and <- calcNoisyAndFrame(list(c("True", "False"), c("True", "False")),
                        c("Right", "Wrong"))
stopifnot (all(and$Right==c(1,0,0,0)))

## DINA, logical-and except that it allows for a small chance of slipping.
dina <- calcNoisyAndFrame(list(c("True", "False"), c("True", "False")),
                        noSlip=.9)
stopifnot (all(abs(dina$Right-c(.9,0,0,0))<.0001))

##NIDA, logical-and except that inputs can randomly be bypassed
nida <- calcNoisyAndFrame(list(c("True", "False"), c("True", "False")),
                        bypass=c(.3,.4))
stopifnot (all(abs(nida$Right-c(1,.3,.4,.12))<.0001))

##Full Noisy And distribution
noisyAnd <- calcNoisyAndFrame(list(c("True", "False"), c("True", "False")),
                        noSlip=.9, bypass=c(.3,.4))

stopifnot (all(abs(noisyAnd$Right-c(1,.3,.4,.12)*.9)<.0001))

thresh <- calcNoisyAndFrame(list(c("H", "M", "L"), c("H", "M", "L")),
                        c("Right", "Wrong"),
                        threshold=c("M", "H"))
stopifnot (all(thresh$Right==c(1,1,0, 0,0,0, 0,0,0)))
```

calcNoisyOrTable

Calculate the conditional probability table for a Noisy-Or distribution

Description

Calculates the conditional probability table for a noisy-and distribution. This follows a logical model where at least one inputs must be true for the output to be true; however, some "noise" is allowed that produces random deviations from the pure logic.

Usage

```
calcNoisyOrTable(skillLevels, obsLevels = c("True", "False"),
                 suppression = rep(0, length(skillLevels)), noGuess = 1,
                 thresholds = sapply(skillLevels, function(states) states[1]))
calcNoisyOrFrame(skillLevels, obsLevels = c("True", "False"),
                 suppression = rep(0, length(skillLevels)), noGuess = 1,
                 thresholds = sapply(skillLevels, function(states) states[1]))
```

Arguments

skillLevels	A list of character vectors giving names of levels for each of the condition variables.
obsLevels	A character vector giving names of levels for the output variables from highest to lowest. As a special case, can also be a vector of integers. Its length should be 2, and the first value is considered to be logically equivalent to "true".
suppression	A vector of the same length as skillLevels. For each parent variable, this represents the probability that the process will act as if that input condition is not met, even if it is met.
noGuess	A scalar value between 0 and 1. This represents the probability that the the output will be false even when all of the inputs are false (e.g., 1-guessing probability).
thresholds	If the input variables have more than two states, values that are equal to or higher than this threshold are considered true. It is assumed that the states of the variables are ordered from highest to lowest.

Details

The noisy-or distribution assumes that both the input and output variables are binary. Basically, the output should be true if any of the inputs are true. Let $S_k = 1$ if the k th input is true, and let q_k be the suppression parameter corresponding to the k th input variable. (If the S_k 's represent a skill, then q_k represents the probability that an examinee who has that skill will fail to correctly apply it.) Then the probability of the true state for the output variable will be:

$$\Pr(X = \text{True} | \mathbf{S}) = 1 - q_0 \prod_k q_k^{1-S_k},$$

where q_0 (the noGuess parameter) is the probability that the output will be false even when all of the inputs are false.

It is assumed that all variables are ordered from highest to lowest state, so that the first state corresponds to "true" the others to false. If the input variable has more than two states, then it can be reduced to a binary variable by using the threshold argument. Any values which are equal to or higher than the threshold for that variable are assumed to be true. (In this case, higher means closer to the the beginning of the list of possible values.)

Value

For calcNoisyOrTable, a matrix whose rows correspond configurations of the parent variable states (skillLevels) and whose columns correspond to obsLevels. Each row of the table is a

probability distribution, so the whole matrix is a conditional probability table. The order of the parent rows is the same as is produced by applying `expand.grid` to `skillLevels`.

For `calcNoisyOrFrame` a data frame with additional columns corresponding to the entries in `skillLevels` giving the parent value for each row.

Note

This is related to the DINO and NIDO models, but uses a slightly different parameterization. In particular, if the `noSlip` parameter is omitted, it is a noisy input deterministic and-gate (NIDO), and if the `bypass` parameters are omitted, it is similar to a deterministic input noisy and-gate (DINO), except it lacks a `slip` parameter.

Author(s)

Russell Almond

References

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (Draft) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Diez, F. J. (1993) Parameter adjustment in Bayes networks. The generalized noisy OR-gate. In Heckerman and Mamdani (eds) *Uncertainty in Artificial Intelligence 93*. Morgan Kaufmann. 99–105.

Srinivas, S. (1993) A generalization of the Noisy-Or model, the generalized noisy OR-gate. In Heckerman and Mamdani (eds) *Uncertainty in Artificial Intelligence 93*. Morgan Kaufmann. 208–215.

See Also

[calcDSTable](#), [calcDNTable](#), [calcDPCTable](#), [expand.grid](#), [calcNoisyOrTable](#)

Examples

```
## Logical or table
or <- calcNoisyOrFrame(list(c("True", "False"), c("True", "False")),
                        c("Right", "Wrong"))
stopifnot (all(or$Right==c(1,1,1,0)))

## DINO, logical-or except that it allows for a small chance of slipping.
dino <- calcNoisyOrFrame(list(c("True", "False"), c("True", "False")),
                        noGuess=.9)
stopifnot (all(abs(dino$True-c(1.0,1.0,1.0,.1))<.0001))

##NIDO, logical-or except that inputs can randomly be bypassed
nido <- calcNoisyOrFrame(list(c("True", "False"), c("True", "False")),
```

```

                                suppression=c(.3,.4))
stopifnot (all(abs(nido$True-c(.88,.6,.7,0))<.0001))

##Full Noisy Or distribution
noisyOr <- calcNoisyOrFrame(list(c("True","False"),c("True","False")),
                             noGuess=.9,suppression=c(.3,.4))

stopifnot (all(abs(noisyOr$False-c(.12,.4,.3,1)*.9)<.0001))

thresh <- calcNoisyOrFrame(list(c("H","M","L"),c("H","M","L")),
                             c("Right","Wrong"),
                             threshold=c("M","H"))
stopifnot (all(thresh$Right==c(1,1,1, 1,1,0, 1,1,0)))

```

colorspread

Produces an ordered palate of colours with the same hue.

Description

This takes a colour specification, and produces an ordered series of colours by manipulating the saturate (and possibly value) of the color, leaving the hue constant. This produces a colour palate suitable for plotting ordered factors, which looks good on a colour display, but also reproduces well on a grayscale printer (or for persons with limited colour perception).

Usage

```
colorspread(col, steps, maxsat = FALSE, rampval = FALSE)
```

Arguments

col	A color in any format suitable as input to col2rgb .
steps	A integer describing the number of colors to create.
maxsat	A logical value. If true, the final color in the series will have saturation 1, instead of whatever is appropriate for the input.
rampval	A logical value. If true, the value as well as the saturation of the color is ramped.

Details

The colour is converted to a RGB value using [col2rgb](#) and then to an HSV value using [rgb2hsv](#). The saturation is then scaled into steps equal intervals. If requested, the value is scaled as well.

Value

A character vectors of length steps giving the colour palate from lowest to highest intensity. This is suitable to passing to the col argument of most graphics functions.

Note

Many of the built-in colours come with 4 intensity variants are meant to work well together. In some cases an expression like `paste("firebrick", 1:4, sep="")` may work better than `colorsread`. To see the built-in colours, use the [colors](#) function.

Author(s)

Russell Almond

See Also

[compareBars](#), `link{stackedBars}`

Examples

```
barplot(rep(1,4),col=colorsread("slategray",4))
barplot(rep(1,4),col=colorsread("slategray",4,maxsat=TRUE))
barplot(rep(1,4),col=colorsread("violetred",4))
barplot(rep(1,4),col=colorsread("violetred",4,rampval=TRUE))
```

compareBars

Produces comparison stacked bar charts for two sets of groups

Description

This produces set of stacked bar charts grouped for comparison between two groups. For example, if suppose that there is a set of probabilities over a collection of proficiency variables measures both before and after obtaining a certain piece of evidence. The `compareBars` function would produce stacked bar charts which compare the prior and posterior probabilities for each variable.

Usage

```
compareBars(data1, data2, profindex,
            groupNames = c(deparse(data1), deparse(data2)), ...,
            ylim = c(min(offsets) - 0.25, max(1 + offsets)),
            cex.names = par("cex.axis"), digits = 2, legend.loc = c(0,1),
            legend.cex = par("cex"), col = par("col"), col1 = NULL,
            col2 = NULL, main = NULL, sub = NULL, xlab = NULL,
            ylab = NULL, rotlab = FALSE)

compareBars2(data1, data2, profindex,
            groupNames=c("Prior","Post"), error.bars=2, scale=100,
            err.col="gray20", ..., ylim = NULL)
```

Arguments

<code>data1</code>	Data set with first (prior) values
<code>data2</code>	Data set with second (posterior) values
<code>profindex</code>	Index of one of the proficiency levels which will be used as the baseline for the stacked bar charts.
<code>groupNames</code>	Names of the groups represented by <code>data1</code> and <code>data2</code> respectively.
<code>...</code>	Other arguments to <code>barplot</code> .
<code>ylim</code>	Default limits for Y axis.
<code>cex.names</code>	Character magnification for names.
<code>digits</code>	Number of digits for overlaid numeric variables.
<code>legend.loc</code>	Location for legend, see legend .
<code>legend.cex</code>	Character magnification for legend.
<code>col</code>	The normal graphics <code>col</code> parameter (see par , passed through to other graphics operators using <code>...</code>). Is also the default for <code>col1</code> and <code>col2</code> if those values are not supplied.
<code>col1</code>	Color scale for the first data set. This should be a vector of colors equal to the number of groups.
<code>col2</code>	Color scale for the second data set. This should be a vector of colors equal to the number of groups.
<code>main</code>	Character scalar giving main title (see title).
<code>sub</code>	Character scalar giving sub title (see title).
<code>xlab</code>	Character scalar giving x-axis label (see title).
<code>ylab</code>	Character scalar giving y-axis label (see title).
<code>rotlab</code>	If TRUE labels are rotated 90 degrees.
<code>error.bars</code>	The number of standard errors for error bars.
<code>err.col</code>	The color for error bars.
<code>scale</code>	Scales data as probabilities (<code>scale=1</code>) or percentages (<code>scale=100</code>).

Note

The function `compareBars2` is a somewhat experimental extension to `compareBars` which adds error bars to the posterior. The result is not entirely satisfactory, and this function may change with future releases.

Author(s)

Russell Almond

References

Almond, R. G., Shute, V. J., Underwood, J. S., and Zapata-Rivera, J.-D (2009). Bayesian Networks: A Teacher's View. *International Journal of Approximate Reasoning*. **50**, 450-460.

See Also

[stackedBars](#), [colspread](#), [buildFactorTab](#), [barplot](#)

Examples

```

margins.prior <- data.frame (
  Trouble=c(Novice=.19,Semester1=.24,Semester2=.28,Semseter3=.20,Semester4=.09),
  NDK=c(Novice=.01,Semester1=.09,Semester2=.35,Semseter3=.41,Semester4=.14),
  Model=c(Novice=.19,Semester1=.28,Semester2=.31,Semseter3=.18,Semester4=.04)
)

margins.post <- data.frame(
  Trouble=c(Novice=.03,Semester1=.15,Semester2=.39,Semseter3=.32,Semester4=.11),
  NDK=c(Novice=.00,Semester1=.03,Semester2=.28,Semseter3=.52,Semester4=.17),
  Model=c(Novice=.10,Semester1=.25,Semester2=.37,Semseter3=.23,Semester4=.05))

foo <-
compareBars(margins.prior,margins.post,3,c("Prior","Post"),
  main="Margins before/after Medium Trouble Shooting Task",
  sub="Observables:  cfgCor=Medium, logCor=High, logEff=Medium",
  legend.loc = "topright",
  cex.names=.75, col1=hsv(h=.1,s=.2*1:5-.1,alpha=1),
  col2=hsv(h=.6,s=.2*1:5-.1,alpha=1))

compareBars2(margins.prior,25*margins.post,3,c("Prior","Post"),
  main="Margins before/after Medium Trouble Shooting Task",
  sub="Observables:  cfgCor=Medium, logCor=High, logEff=Medium",
  legend.loc = "topright",
  cex.names=.75, col1=hsv(h=.1,s=.2*1:5-.1,alpha=1),
  col2=hsv(h=.6,s=.2*1:5-.1,alpha=1))

```

Compensatory

DiBello–Samejima combination function

Description

These functions take a vector of “effective theta” values for a collection of parent variables and calculates the effective theta value for the child variable according to the named rule. Used in calculating DiBello–Samejima and DiBello–Normal probability tables. These all have one slope parameter (alpha) per parent variable.

Usage

```

Compensatory(theta, alphas, beta)
Conjunctive(theta, alphas, beta)
Disjunctive(theta, alphas, beta)

```

Arguments

theta	A matrix of effective theta values whose columns correspond to parent variables and whose rows correspond to possible skill profiles.
alphas	A vector of discrimination parameters in the same order as the columns of theta. (Note these function expect discrimination parameters and not log discrimination parameters as used in <code>calcDSTable</code> .)
beta	A difficulty (-intercept) parameter.

Details

For *Compensatory*, the combination function for each row is:

$$(alphas[1] * theta[1] + \dots + alphas[K] * theta[K]) / \sqrt{K} - beta$$

where K is the number of parents. (The \sqrt{K} is a variance stabilization parameter.)

For *Conjunctive*, the combination function for each row is:

$$\min(alphas[1] * theta[1], \dots, alphas[K] * theta[K]) - beta$$

For *Disjunctive*, the combination function for each row is:

$$\max(alphas[1] * theta[1], \dots, alphas[K] * theta[K]) - beta$$

Value

A vector of normal deviates corresponding to the effective theta value. Length is the number of rows of thetas.

Note

These functions expect the unlogged discrimination parameters, while `calcDSTable` expect the log of the discrimination parameters. The rationale is that log discrimination is bound away from zero, and hence a more natural space for MCMC algorithms. However, it is poor programming design, as it is liable to catch the unwary.

These functions are meant to be used as structure functions in the DiBello–Samejima and DiBello–Normal models. Other structure functions are possible and can be excepted by those functions as long as they have the same signature as these functions.

Author(s)

Russell Almond

References

- Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (Draft) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.
- Almond, R.G., DiBello, L., Jenkins, F., Mislevy, R.J., Senturk, D., Steinberg, L.S. and Yan, D. (2001) Models for Conditional Probability Tables in Educational Assessment. *Artificial Intelligence and Statistics 2001* Jaakkola and Richardson (eds), Morgan Kaufmann, 137–143.

See Also

[effectiveThetas](#), [calcDSTable](#), [calcDNTable](#), [calcDPCTable](#), [OffsetConjunctive](#), [eThetaFrame](#)

Examples

```
thetas <- expand.grid(list(S1=(3:1 -2), S2 = (3:1 -2)))
Compensatory(thetas, c(S1=1.25,S2=.75), 0.33)
Conjunctive(thetas, c(S1=1.25,S2=.75), 0.33)
Disjunctive(thetas, c(S1=1.25,S2=.75), 0.33)

skill <- c("High","Medium","Low")
eThetaFrame(list(S1=skill,S2=skill), c(S1=1.25,S2=.75), 0.33, "Compensatory")
eThetaFrame(list(S1=skill,S2=skill), c(S1=1.25,S2=.75), 0.33, "Conjunctive")
eThetaFrame(list(S1=skill,S2=skill), c(S1=1.25,S2=.75), 0.33, "Disjunctive")
```

dataTable	<i>Constructs a table of counts from a set of discrete observations.</i>
-----------	--

Description

This constructs a table of counts in a special format useful for conditional probability tables. The rows correspond to configurations of the parent variables and the columns correspond to possible states of the child variables.

Usage

```
dataTable(data, parents, child, childStates)
```

Arguments

data	A data frame whose columns contain variables corresponding to parent and child.
parents	A vector of names for the columns in data corresponding to the parent variables.
child	The name of the child variable, should refer to a column in data.
childStates	A character vector giving names of levels for the output variables from highest to lowest.

Details

Apply the function [table](#) to generate a table of counts for the indicated variables (subsetting the table if necessary). Then reformats this into a matrix whose columns correspond to the child variable.

Value

A matrix whose columns correspond to childStates and whose rows correspond to the possible combinations of parents.

Author(s)

Russell Almond

See Also[table](#), [calcDSlike](#)**Examples**

```

skill1l <- c("High", "Medium", "Low")
skill3l <- c("High", "Better", "Medium", "Worse", "Low")
correctL <- c("Correct", "Incorrect")

x <- read.csv(paste(library(help="CPTtools")$path,
  "testFiles", "randomPinned100.csv",
  sep=.Platform$file.sep),
  header=FALSE, as.is=TRUE,
  col.names = c("Skill1", "Skill2", "Skill3",
    "Comp.Correct", "Comp.Grade",
    "Conj.Correct", "Conj.Grade",
    "Cor.Correct", "Cor.Grade",
    "Dis.Correct", "Dis.Grade",
    "Inhib.Correct", "Inhib.Grade"
  ))
x[, "Skill1"] <- ordered(x[, "Skill1"], skill1l)
x[, "Skill3"] <- ordered(x[, "Skill3"], skill3l)
x[, "Comp.Correct"] <- ordered(x[, "Comp.Correct"], correctL)

tab <- dataTable(x, c("Skill1", "Skill3"), "Comp.Correct", correctL)
data.frame(expand.grid(list(Skill1=skill1l, Skill3=skill3l)), tab)

```

effectiveThetas

*Assigns effective theta levels for categorical variable***Description**

Calculates a vector of normal quantiles corresponding to effective theta levels for a categorical variable for use in a DiBello-Samejima distribution.

Usage

```
effectiveThetas(nlevels)
```

Arguments

nlevels Integer giving the number of levels of the categorical variable.

Details

The DiBello–Samejima models map levels of categorical values into effective “theta” values, or corresponding continuous values. These can then be input into IRT equations to calculate cell probabilities.

The default algorithm works by assuming that the categories are created by cutting the normal distribution into equal probability intervals. The value returned for each interval is the midpoint (wrt the normal measure) of that interval.

Value

A vector of normal quantiles of length `nlevels`.

Author(s)

Russell Almond

References

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (Draft) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.

Examples

```
effectiveThetas(5)
```

eThetaFrame	<i>Constructs a data frame showing the effective thetas for each parent combination.</i>
-------------	--

Description

This evaluates the combination function but not the link function of an effective theta distribution. It produces a table of effective thetas one for each configuration of the parent values according to the combination function given in the `model` argument.

Usage

```
eThetaFrame(skillLevels, lnAlphas, beta, rule = "Compensatory")
```

Arguments

<code>skillLevels</code>	A list of character vectors giving names of levels for each of the condition variables.
<code>lnAlphas</code>	A vector of log slope parameters. Its length should be either 1 or the length of <code>skillLevels</code> , depending on the choice of rule.
<code>beta</code>	A vector of difficulty (-intercept) parameters. Its length should be either 1 or the length of <code>skillLevels</code> , depending on the choice of rule.
<code>rule</code>	Function for computing effective theta (see Details).

Details

The DiBello framework for creating conditional probability tables for Bayesian network models using IRT-like parameters unfolds in three steps.

1. Each level of each input variable is assigned an “effective theta” value — a normal value to be used in calculations.
2. For each possible skill profile (combination of states of the parent variables) the effective thetas are combined using a combination function. This produces an “effective theta” for that skill profile. The function rule determines the rule for combination.
3. The effective theta is input into a link function (e.g., Samejima’s graded-response function) to produce a probability distribution over the states of the outcome variables.

This function applies the first two of those steps and returns a data frame with the original skill levels and the effective thetas.

The parent (conditioning) variables are described by the `skillLevels` argument which should provide for each parent variable in order the names of the states ranked from highest to lowest value. The original method (Almond et al., 2001) used equally spaced points on the interval $[-1, 1]$ for the effective thetas of the parent variables. The current implementation uses the function `effectiveThetas` to calculate equally spaced points on the normal curve.

The combination of the individual effective theta values into a joint value for effective theta is done by the function reference by rule. This should be a function of three arguments: `theta` — the vector of effective theta values for each parent, `alphas` — the vector of discrimination parameters, and `beta` — a scalar value giving the difficulty. The initial distribution supplies five functions appropriate for use with `calcDSTable`: `Compensatory`, `Conjunctive`, and `Disjunctive`, `OffsetConjunctive`, and `OffsetDisjunctive`. The last two have a slightly different parameterization: `alpha` is assumed to be a scalar and `betas` parameter is vector valued. Note that the discrimination and difficulty parameters are built into the structure function and not the IRT curve.

Value

For a data frame with one column for each parent variable and an additional column for the effective theta values. The number of rows is the product of the number of states in each of the components of the `skillLevels` argument.

Author(s)

Russell Almond

References

- Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (Draft) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.
- Almond, R.G., DiBello, L., Jenkins, F., Mislevy, R.J., Senturk, D., Steinberg, L.S. and Yan, D. (2001) Models for Conditional Probability Tables in Educational Assessment. *Artificial Intelligence and Statistics 2001* Jaakkola and Richardson (eds), Morgan Kaufmann, 137–143.

See Also

[effectiveThetas](#), [Compensatory](#), [OffsetConjunctive](#), [calcDNTTable](#), [calcDSTable](#), [calcDPCTable](#), [expand.grid](#)

Examples

```
skill <- c("High", "Medium", "Low")
eThetaFrame(list(S1=skill, S2=skill), log(c(S1=1.25, S2=.75)), 0.33,
             "Compensatory")
eThetaFrame(list(S1=skill, S2=skill), log(c(S1=1.25, S2=.75)), 0.33,
             "Conjunctive")
eThetaFrame(list(S1=skill, S2=skill), log(c(S1=1.25, S2=.75)), 0.33,
             "Disjunctive")
eThetaFrame(list(S1=skill, S2=skill), log(1.0), c(S1=0.25, S2=-0.25),
             "OffsetConjunctive")
eThetaFrame(list(S1=skill, S2=skill), log(1.0), c(S1=0.25, S2=-0.25),
             "OffsetDisjunctive")
```

getTableStates

Gets meta data about a conditional probability table.

Description

Fetches the names of the parent variables, or the names of the states of the child variable from a conditional probability table.

Usage

```
getTableStates(table)
```

Arguments

`table` A conditional probability table expressed as a data frame.

Details

These functions assume that `table` is a conditional probability table (or a set of hyper-Dirichlet parameters) which is shaped like a data frame. Columns in the data frame which are factors are assumed to hold values for the parent (conditioning) variables. Columns in the data frame which are numeric are assumed to correspond to possible states of the child (dependent) variable.

Value

For `getTableParents()`, a character vector giving the names of the parent variables (factor columns).

For `getTableStates()`, a character vector giving the names of the child states (numeric columns).

Note

StatShop usually assumes that the states are ordered from the *highest* to the *lowest* possible values, for example ‘High’, ‘Med’, ‘Low’.

Author(s)

Russell Almond

See Also

[rescaleTable](#)

Examples

```
#conditional table
X2.ptf <- data.frame(Theta=c("Expert","Novice"),
                    correct=c(4,2),
                    incorrect=c(2,4))

#Unconditional table
Theta.ptf <- data.frame(Expert=3,Novice=3)

stopifnot(
  identical(getTableStates(X2.ptf), c("correct","incorrect")),
  identical(getTableStates(Theta.ptf), c("Expert","Novice")),
  identical(getTableParents(X2.ptf), "Theta"),
  identical(getTableParents(Theta.ptf), character(0))
)
```

gradedResponse	<i>A link function based on Samejima’s graded response</i>
----------------	--

Description

This function converts a matrix of effective theta values into a conditional probability table by applying Samejima’s graded response model to each row of the table.

Usage

```
gradedResponse(et, linkScale = NULL, obsLevels = NULL)
```

Arguments

et	A matrix of effective theta values. There should be one row in this table for each configuration of the parent variables of the conditional probability table and one column for each state of the child variables except for the last.
linkScale	Unused. For compatibility with other link functions.
obsLevels	A character vector giving the names of the child variable states. If supplied, it should have length ncol(et)+1.

Details

This function takes care of the third step in the algorithm of [calcDPCTable](#). Its input is a matrix of effective theta values (comparable to the last column of the output of [eThetaFrame](#)), one column for each of the child variable states (`obsLevels`) except for the last one. Each row represents a different configuration of the parent variables. The output is the conditional probability table.

Let X be the child variable of the distribution, and assume that it can take on M possible states labeled x_1 through x_M in increasing order. The graded response model defines a set of functions $Z_m(\theta_k)$ for $m = 2, \dots, M$, where

$$Pr(X \geq x_m | \theta_k) = \text{logit}^{-1} - D * Z_m(\theta_k)$$

The conditional probabilities for each child state given the effective thetas for the parent variables is then given by

$$Pr(X = x_m | \theta_k) = \frac{\sum_{r=1}^m Z_r(\theta_k)}{\sum_{r=1}^M Z_r(\theta_k)}$$

The $K \times M - 1$ matrix `et` is the values of $Z_m(\theta_k)$. This function then performs the rest of the generalized partial credit model. This is a generalization of Muraki (1992), because the functions $Z_m(\cdot)$ are not restricted to be the same functional form for all m .

If supplied `obsLevels` is used for the column names.

Value

A matrix with one more column than `et` giving the conditional probabilities for each configuration of the parent variables (which correspond to the rows).

Note

The `linkScale` parameter is unused. It is for compatibility with other link function choices.

Author(s)

Russell Almond

References

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (Draft) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, **16**, 159-176. DOI: 10.1177/014662169201600206

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, **34**, (No. 4, Part 2).

I also have planned a manuscript that describes these functions in more detail.

See Also

[eThetaFrame](#), [calcDPCTable](#), [gradedResponse](#),

Examples

```
## Set up variables
skill1l <- c("High", "Medium", "Low")
correctL <- c("Correct", "Incorrect")
pcreditL <- c("Full", "Partial", "None")
gradeL <- c("A", "B", "C", "D", "E")

## Get some effective theta values.
et <- effectiveThetas(3)

gradedResponse(matrix(et, ncol=1), NULL, correctL)

gradedResponse(outer(et, c(Full=1, Partial=-1)), NULL, pcreditL)

gradedResponse(outer(et, c(A=2, B=1, C=0, D=-1)), NULL, gradeL)
```

localDepTest	<i>Tests for conditional independence between two variables given a third</i>
--------------	---

Description

The function `ciTest` takes a 3-way contingency table and tests for the conditional independence of the first two variables given the third. The function `localDepTest` is a wrapper which builds the table from factor variables. In psychometrics, when the third variable represents a latent proficiency and the first two item responses, this is sometimes called *local independence*.

Usage

```
localDepTest(obs1, obs2, prof)
ciTest(tab)
```

Arguments

<code>obs1</code>	A factor variable representing the first observable outcome.
<code>obs2</code>	A factor variable representing the second observable outcome.
<code>prof</code>	A factor variable representing the proficiency level, or any variable that is thought to render <code>obs1</code> and <code>obs2</code> independent.
<code>tab</code>	A three-way table (see table) where the first two dimensions represent the observable variables and the third the proficiency variable.

Details

Let 1 and 2 represent obs1 and obs2 respectively and let 3 represent prof. In the case of `ciTest`, 1, 2 and 3 represent the first second and 3rd dimensions of `tab`. These function then compare the undirected model [13][23] (1 and 2 are conditionally independent given 3) to the unrestricted model [123]. The result is a chi-square statistic comparing the two models, high values of the chi-square statistic indicate a better fit of the unrestricted model compared to the conditional independence model.

Note that the Cochran-Mantel-Haenszel statistic (see [mantelhaen.test](#)) is similar, but it assumes that there is no three-way interaction, so it essentially tests [13][23] versus [12][13][23].

Value

A list with three elements:

G2	The chi-square comparison between the two models.
df	The degrees of freedom for the test.
p	The percentage point for G2 in a central chi-square distribution with df degrees of freedom, i.e., the p-value.

Author(s)

Russell Almond

References

Bishop, Feinberg and Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

See Also

[buildFactorTab](#), [mantelhaen.test](#), [UCBAdmissions](#)

Examples

```
data(UCBAdmissions)
ciTest(UCBAdmissions)

data(ACED)
table(ACED.items$tCommonRatio1a.1, ACED.items$tCommonRatio2a.1,
      cut(ACED.items$Correct, 3))
localDepTest(ACED.items$tCommonRatio1a.1, ACED.items$tCommonRatio2a.1,
             cut(ACED.items$Correct, 3))
```

mapDPC

*Finds an MAP estimate for a discrete partial credit CPT***Description**

This finds a set of parameters for a given discrete partial credit model which produces the best fit to the first argument. It is assumed that the first argument is a table produced by adding observed counts to a prior conditional probability table. Thus the result is a maximum a posterior (MAP) estimate of the CPT. The parametric structure of the table is given by the rules and link parameters.

Usage

```
mapDPC(postTable, skillLevels, obsLevels, lnAlphas, betas,
       rules = "Compensatory", link = "partialCredit",
       linkScale=NULL, Q=TRUE, tvals=lapply(skillLevels,
       function (sl) effectiveThetas(length(sl))),
       ...)
```

Arguments

postTable	A table of cell counts which should have the same structure as the output of calcDPCTable (skillLevels, obsLevels, lnAlphas, betas, rules, link, linkScale). As zero counts would cause problems, the prior conditional probability table is normally added to the counts to make the argument a posterior counts.
skillLevels	A list of character vectors giving names of levels for each of the condition variables.
obsLevels	A character vector giving names of levels for the output variables from highest to lowest. As a special case, can also be a vector of integers.
lnAlphas	A list of vectors of initial values for the log slope parameters. Its length should be 1 or length(obsLevels)-1. The required length of the individual component vectors depends on the choice of rule (and is usually either 1 or the length of skillLevels).
betas	A list of vectors of initial values for the difficulty (-intercept) parameters. Its length should be 1 or length(obsLevels)-1. The required length of the individual component vectors depends on the choice of rule (and is usually either 1 or the length of skillLevels).
rules	A list of functions for computing effective theta (see Details). Its length should be length(obsLevels)-1 or 1 (implying that the same rule is applied for every gap.)
link	The function that converts a table of effective thetas to probabilities
linkScale	Initial values for the optional scale parameter for the link function. This is only used with certain choices of link function.
Q	This should be a Q matrix indicating which parent variables are relevant for which state transitions. It should be a number of states minus one by number of parents logical matrix. As a special case, if all variable are used for all levels, then it can be a scalar value.

tvals	A list of the same length as skillLevels. Each element should be a numeric vector values on the theta (logistic) scale corresponding to the levels for that parent variable. The default spaces them equally according to the normal distribution (see effectiveThetas).
...	Additional arguments passed to the optim function.

Details

The purpose of this function is to try to estimate the values of a discrete partial credit model. The structure of the model is given by the rules and link arguments: the form of the table produces is like the output of [calcDPCTable](#)(skillLevels, obsLevels, lnAlphas, betas, rules, link, linkScale). It tries to find the values of lnAlphas and betas (and if used linkScale) parameters which are most likely to have generated the data in the postTable argument. The lnAlphas, betas and linkScale arguments provide the initial values for those parameters.

Let $p_{i,j}$ be the value in the i th row and the j th column of the conditional probability table output from [calcDPCTable](#)(skillLevels, obsLevels, lnAlphas, betas, rules, link, linkScale), and let $x_{i,j}$ be the corresponding elements of postTable. The mapDPC function uses [optim](#) to find the value of the parameters that minimizes the deviance,

$$-2 * \sum_i \sum_j x_{i,j} \log(p_{i,j}).$$

Value

A list with components:

lnAlphas A vector of the same structure as lnAlphas containing the estimated values.

betas A veto of the same structure as betas containing the estimated values.

linkScale If the linkScale was supplied, the estimated value.

convergence An integer code. 0 indicates successful completion, positive values are various error codes (see [optim](#)).

value The deviance of the fit DPC model.

The list is the output of the [optim](#) function, which has other components in the output. See the documentation of that function for details.

Author(s)

Russell Almond

References

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (Draft) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, **16**, 159-176. DOI: 10.1177/014662169201600206

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, **34**, (No. 4, Part 2).

I also have planned a manuscript that describes these functions in more detail.

See Also

[optim](#)), [calcDPCTable](#), [Compensatory](#), [OffsetConjunctive](#), [gradedResponse](#), [partialCredit](#)

Examples

```
pLevels <- list(Skill1=c("High","Med","Low"))
obsLevels <- c("Full","Partial","None")

trueLnAlphas <- list(log(1),log(.25))
trueBetas <- list(2, -.5)

priorLnAlphas <- list(log(.5),log(.5))
priorBetas <- list(1, -1)

truedist <- calcDPCTable(pLevels,obsLevels,trueLnAlphas,trueBetas,
                        rules="Compensatory",link="partialCredit")
prior <- calcDPCTable(pLevels,obsLevels,priorLnAlphas,priorBetas,
                     rules="Compensatory",link="partialCredit")

post1 <- prior + round(truedist*1000)

map1 <- mapDPC(post1,pLevels,obsLevels,priorLnAlphas,priorBetas,
              rules="Compensatory",link="partialCredit")

if (map1$convergence != 0) {
  warning("Optimization did not converge:", map1$message)
}

postLnAlphas <- map1$lnAlphas
postBetas <- map1$betas
fitdist <- calcDPCTable(pLevels,obsLevels,map1$lnAlphas,map1$betas,
                      rules="Compensatory",link="partialCredit")
## Tolerance for recovery test.
tol <- .01
maxdistdif <- max(abs(fitdist-truedist))
if (maxdistdif > tol) {
  stop("Posterior and True CPT differ, maximum difference ",maxdistdif)
}
if (any(abs(unlist(postLnAlphas)-unlist(trueLnAlphas))>tol)) {
  stop("Log(alphas) differ by more than tolerance")
}
if (any(abs(unlist(postBetas)-unlist(trueBetas))>tol)) {
  stop("Betas differ by more than tolerance")
}
```


Description

Summary statistics for a data set consisting of marks on five mathematics exams originally presented in Mardia, Kent and Bibby (1979). The raw data are not given, but rather the summary statistics reported in Whittaker (1990) are provided.

Usage

```
data(MathGrades)
```

Format

A list consisting of the following components:

varnames Names of the five tests.

means Average score on each test.

var Covariance matrix.

cor Correlation matrix.

icdiag Diagonal of the inverse covariance matrix.

pcor Partial correlation (scaled inverse correlation) matrix.

pvecs A set of marginal distributions for discrete variables corresponding to the five assessments.

Source

Summary statistics reported here are taken from Whittaker (1990).

Original data on 88 students is reported in Mardia, Kent and Bibby (1979).

References

Mardia, K.V. and Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis*. Academic Press.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

Examples

```
data(MathGrades)
```

```
##Note: Some of these tests may return false due to machine precision
##issues.
round(scaleMatrix(MathGrades$var),2) == MathGrades$cor
round(diag(solve(MathGrades$cor)),2) == MathGrades$icdiag
round(scaleMatrix(solve(MathGrades$cor)),2) == MathGrades$pcor
```

mcSearch

Orders variables using Maximum Cardinality search

Description

Takes a graph described by an incidence matrix, and creates an ordering of the nodes using maximum cardinality search (Tarjan and Yannakakis, 1984). A primary application of this method is to choose an ordering of nodes in an undirected graph to make a corresponding directed graph.

Usage

```
mcSearch(sm, start = colnames(sm)[1])
```

Arguments

sm	A logical matrix whose rows and columns correspond to nodes (variables) and a true value indicates an edge between the variables.
start	The name of the first element.

Details

The sm argument should be an incidence matrix for a graph, with row and column names set to the names of the nodes/variables.

The function returns an ordering of the nodes where each node is chosen so that it has a maximum number of neighbors among those nodes higher in the order.

Ties are broken by choosing the first node in the graph (using the order of the columns) matching the criteria. One special case is the first node which is always an arbitrary choice. The start argument can be used to force a particular selection.

Value

A vector of length equal to the number of rows whose values correspond to the order of the variable in the final order.

Note

If the graph is triangulated, then the ordering produced by mcSearch should be *perfect* — for each node in the order, the set of neighbors of that node which precede it in the ordering is completely connected. Perfect node orderings are useful in going from undirected to directed graphical representations. If we take an undirected graph and a perfect ordering, we define as the parents of a node all of its neighbors which are previous in the order and define as its children all of the nodes which are later in the order, then when converting the graph back to the undirected form no additional “moralization” edges will be required. Thus, this function can be used to generate orders for [buildParentList](#).

Graphical models generally exist only over triangulated graphs. Therefore, and incidence matrix which has been produced through the use of the `structMatrix` function should always work. Tarjan and Yannakakis (1984) prove that the maximum cardinality search always produces a perfect ordering when the graph is triangulated.

When the graph is not triangulated, the maximum cardinality search algorithm can be used to generate “fill-ins” to triangulate the graph. Lauritzen and Spiegelhalter (1988) note this use. While maximum cardinality search will produce an ordering quickly, the ordering itself has now particular optimality properties as far as the triangulated graph which it creates (Almond, 1995).

Author(s)

Russell Almond

References

- Almond, R.G. (1995). *Graphical Belief Modeling*. Chapman and Hall.
- Lauritzen, S.L. and D.J. Spiegelhalter (1988). Local Computation with Probabilities on Graphical Structures and their Application to Expert Systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **50**, 205-247.
- Tarjan, R.E. and M. Yannakakis (1984). Simple Linear-Time Algorithms to test Chordality of Graphs, Test Acyclicity of Hypergraphs, and Selectively Reduce Acyclic Hypergraphs. *Siam J. Comput.* **13**, 566-579.

See Also

`structMatrix`, `buildParentList`

Examples

```
data(MathGrades)
MG.struct <- structMatrix(MathGrades$var)

ord <- mcSearch(MG.struct) # Arbitrary start
orda <- mcSearch(MG.struct, "Algebra") # Put algebra first.

names(sort(orda)) # names of the variables in the chosen order.

# Sort rows and columns of structure matrix by MC order
MG.struct[order(orda), order(orda)]
```

mutualInformation	<i>Calculates Mutual Information for a two-way table.</i>
-------------------	---

Description

Calculates the mutual information for a two-way table of observed counts or a joint probability distribution. The mutual information is a measure of association between two random variables.

Usage

```
mutualInformation(table)
```

Arguments

table	A two way table or probability distribution. Possibly the output of the table command.
-------	--

Details

This is basically the Kullback-Leibler distance between the joint probability distribution and the probability distribution created by assuming the marginal distributions are independent. This is given in the following formula:

$$I[X;Y] = \sum_x \sum_y y \Pr(X = x, Y = y) \log \frac{\Pr(X = x, Y = y)}{\Pr(X = x) \Pr(Y = y)}$$

Author(s)

Russell Almond

References

<http://planetmath.org/encyclopedia/MutualInformation.html>

Shannon (1948) "A Mathematical Theory of Communication."

See Also

[table](#)

Examples

```
## UCBAmissions is a three way table, so we need to
## make it a two way table.
mutualInformation(apply(UCBAmissions,c(1,2),sum))
apply(UCBAmissions,3,mutualInformation)
apply(UCBAmissions,2,mutualInformation)
apply(UCBAmissions,1,mutualInformation)
```

`numericPart`*Splits a mixed data frame into a numeric matrix and a factor part.*

Description

The function `numericPart()` converts a `data.frame` to a matrix, by dropping columns which contain non-numeric data. The function `factorPart` grabs the state information by selecting only columns which are factors.

Usage

```
numericPart(table)
```

Arguments

`table` A `data.frame` object.

Details

The primary purpose is to split a conditional probability distribution in data frame format (with a set of factor rows identifying the states of the parent variables) to a table of just the numbers, and a data frame of just the factors so that they can be tackled separately.

Value

A matrix containing just the numeric columns of the data frame.

Author(s)

Russell Almond

See Also

[data.frame](#), [matrix](#), [data.matrix](#)

Examples

```
name <-c("Shahrazad", "Marguerite")
height <- c(34, 36)
weight <- c(28, 26)
twins <- data.frame(name=I(name),height=height, weight=weight)
numericPart(twins)
```

Description

The observable characteristic plot is an analogue of the item characteristic curve for latent class models. Estimates of the class success probability for each latent is plotted. Reference lines are added for various groups which are expected to have roughly the same probability of success.

Usage

```
OCP(x, n, lclabs, pi, pilab = names(pi), lcnames = names(x),
    a = 0.5, b = 0.5, reflty = 1, ..., newplot = TRUE,
    main = NULL, sub = NULL, xlab = "Latent Classes", ylab = "Probability",
    cex = par("cex"), xlim = NULL, ylim = NULL, cex.axis = par("cex.axis"))
OCP2(x, n, lclabs, pi, pilab = names(pi), lcnames = names(x),
    set1 = seq(1, length(x)-1, 2), setlabs = c("Set1", "Set2"),
    setat = -1, a = 0.5, b = 0.5, reflty = 1, ..., newplot = TRUE,
    main = NULL, sub = NULL, xlab = "Latent Classes", ylab = "Probability",
    cex = par("cex"), xlim = NULL, ylim = NULL, cex.axis = par("cex.axis"))
```

Arguments

x	Vector of success counts for each latent class.
n	Vector of of the same size as x of latent class sizes.
lclabs	Character vector of plotting symbols for each latent class
pi	Vector of probability estimates for various levels.
pilab	Character vector of names for each level.
lcnames	Character vector of names for each latent class for axis.
set1	For OCP2, a vector of indexes of the elements of x that form the first set.
setlabs	Character vectors of set labels for OCP2.
setat	Numeric scalar giving the x-coordinate for set labels.
a	The first parameter for beta pseudo-prior, passed to betaci . This should either be a scalar or the same length as x.
b	The second parameter for beta pseudo-prior, passed to betaci . This should either be a scalar or the same length as n.
reflty	Provides the line type (see par(lty)) for reference lines. Should be scalar or same length as pi.
...	Additional graphics parameters passed to plot.window and title .
newplot	Logical. If true, a new plotting window is created. Otherwise, information is added to existing plotting window.
main	Character scalar giving main title (see title).

sub	Character scalar giving sub title (see title).
xlab	Character scalar giving x-axis label (see title).
ylab	Character scalar giving y-axis label (see title).
cex	Scalar giving character expansion for plotting symbols (see par (cex)).
xlim	Plotting limits for x-axis (see plot.window).
ylim	Plotting limits for y-axis (see plot.window).
cex.axis	Scalar giving character expansion for latent class names (axis tick labels; cex values passed to par (axis)).

Details

Most cognitively diagnostic models for assessments assume that students with different patterns of skills will have different patterns of success and failures. The goal of this plot type is to empirically check to see if the various groups conform to their expected probability.

The assumption is that the population is divided into a number of *latent classes* (or groups of latent classes) and that the class label for each member of the class is known. The latent classes are partitioned into *levels*, where each level is assumed to have roughly the same proportion of success. (This is often $c("-", "+")$ for negative and positive skill patterns, but there could be multiple levels.)

The key idea of the observable characteristic plot is to compare a credibility interval for the success probability in each latent class, to the modelled success probability given by its level. A beta credibility interval is computed for each latent class using [betaci](#)(x, n, a, b), with the expected value set at $(x + a)/(n + a + b)$. The plotting symbols given in [lclabs](#) are plotted at the mean (usually, these correspond to the latent class is in), with vertical error bars determined by [betaci](#). Horizontal reference lines are added at the values given by [pi](#). The idea is that if the error bars cross the reference line for the appropriate level, then the latent class fits the model, if not it does not.

The variant OCP2 is the same except that the latent classes are further partitioned into two sets, and the labels for the latent classes for different sets are plotted on different lines. This is particularly useful for testing whether attributes which are not thought to be relevant to a given problem are actually irrelevant.

Value

The output of [betaci](#) is returned invisibly.

Author(s)

Russell Almond

References

- Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (2015) *Bayesian Networks in Educational Assessment*. Springer. Chapter 10.
- Sinharay, S. and Almond, R.G. (2006). Assessing Fit of Cognitively Diagnostic Models: A case study. *Educational and Psychological Measurement*. **67**(2), 239–257.

Sinharay, S., Almond, R. G. and Yan, D. (2004). Assessing fit of models with discrete proficiency variables in educational assessment. ETS Research Report. <http://www.ets.org/research/researcher/RR-04-07.html>

See Also

[betaci](#)

Examples

```
nn <- c(30,15,20,35)
pi <- c("+=".15, "-=".85)
grouplabs <- c(rep("-",3),"+")
x <- c("(0,0)"=7,"(0,1)"=4,"(1,0)"=2,"(1,1)"=31)
OCP (x,nn,grouplabs,pi,c("-", "+"),ylim=c(0,1), reflty=c(2,4),
    main="Data that fit the model")

x1 <- c("(0,0)"=7,"(0,1)"=4,"(1,0)"=11,"(1,1)"=31)
OCP (x1,nn,grouplabs,pi,c("-", "+"),ylim=c(0,1), reflty=c(2,4),
    main="Data that don't fit the model")

nnn <- c("(0,0,0)"=20,"(0,0,1)"=10,
        "(0,1,0)"=10,"(0,1,1)"=5,
        "(1,0,0)"=10,"(1,0,1)"=10,
        "(1,1,1)"=10,"(1,1,1)"=25)
xx <- c("(0,0,0)"=5,"(0,0,1)"=2,
        "(0,1,0)"=2,"(0,1,1)"=2,
        "(1,0,0)"=2,"(1,0,1)"=0,
        "(1,1,0)"=9,"(1,1,1)"=21)
grouplabs1 <- rep(grouplabs,each=2)

OCP2 (xx,nnn,grouplabs1,pi,c("-", "+"),ylim=c(0,1), reflty=c(2,4),
    setlabs=c("Low Skill3","High Skill3"),setat=-.8,
    main="Data for which Skill 3 is irrelevant")

xx1 <- c("(0,0,0)"=2,"(0,0,1)"=5,
        "(0,1,0)"=1,"(0,1,1)"=3,
        "(1,0,0)"=0,"(1,0,1)"=2,
        "(1,1,0)"=5,"(1,1,1)"=24)
OCP2 (xx1,nnn,grouplabs1,pi,c("-", "+"),ylim=c(0,1), reflty=c(2,4),
    setlabs=c("Low Skill3","High Skill3"),setat=-.8,
    main="Data for which Skill 3 is relevant")
```


Description

These functions take a vector of “effective theta” values for a collection of parent variables and calculates the effective theta value for the child variable according to the named rule. Used in calculating DiBello–Samejima and DiBello–Normal probability tables. These versions have a single slope parameter (alpha) and one difficulty parameter per parent variable.

Usage

```
OffsetConjunctive(theta, alpha, betas)
OffsetDisjunctive(theta, alpha, betas)
```

Arguments

theta	A matrix of effective theta values whose columns correspond to parent variables and whose rows correspond to possible skill profiles.
alpha	A single common discrimination parameter. (Note these function expect discrimination parameters and not log discrimination parameters as used in calcDSTable .)
betas	A vector of difficulty (-intercept) parameters. Its length should be the same as the number of columns in theta.

Details

For `OffsetConjunctive`, the combination function for each row is:

$$\alpha * \min(\theta[1] - \beta[1], \dots, \theta[K] - \beta[K])$$

For `OffsetDisjunctive`, the combination function for each row is:

$$\alpha * \max(\theta[1] - \beta[1], \dots, \theta[K] - \beta[K])$$

Value

A vector of normal deviates corresponding to the effective theta value. Length is the number of rows of thetas.

Note

These functions expect the unlogged discrimination parameters, while `calcDSTable` expect the log of the discrimination parameters. The rationale is that log discrimination is bound away from zero, and hence a more natural space for MCMC algorithms. However, it is poor programming design, as it is liable to catch the unwary.

These functions are meant to be used as structure functions in the DiBello–Samejima and DiBello–Normal models. Other structure functions are possible and can be expected by those functions as long as they have the same signature as these functions.

Author(s)

Russell Almond

References

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (Draft) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.

Almond, R.G., DiBello, L., Jenkins, F., Mislevy, R.J., Senturk, D., Steinberg, L.S. and Yan, D. (2001) Models for Conditional Probability Tables in Educational Assessment. *Artificial Intelligence and Statistics 2001* Jaakkola and Richardson (eds), Morgan Kaufmann, 137–143.

See Also

[effectiveThetas](#), [calcDSTable](#), [calcDNTable](#), [calcDPCTable](#), [Compensatory](#), [eThetaFrame](#)

Examples

```
skill <- c("High", "Medium", "Low")
thetas <- expand.grid(list(S1=(3:1 -2), S2 = (3:1 -2)))
OffsetDisjunctive(thetas, 1.0, c(S1=0.25, S2=-0.25))
OffsetConjunctive(thetas, 1.0, c(S1=0.25, S2=-0.25))
eThetaFrame(list(S1=skill, S2=skill), 1.0, c(S1=0.25, S2=-0.25),
             "OffsetConjunctive")
eThetaFrame(list(S1=skill, S2=skill), 1.0, c(S1=0.25, S2=-0.25),
             "OffsetDisjunctive")
```

parseProbVec

Parses Probability Vector Strings

Description

This takes a bunch of strings of the form "[High: .3, Med: .5, Low: .2]" and parses it into a vector `c(High=.3, Med=.5, Low=.2)`.

Usage

```
parseProbVec(pVec)
parseProbVecRow(splitrow)
```

Arguments

<code>pVec</code>	A string of the form "[High: .3, Med: .5, Low: .2]"
<code>splitrow</code>	A collection of strings "High: .3", "Med: .5", "Low: .2".

Details

StatShop outputs marginal distributions in the format `[state0: val0, state1: val1, ...]`. This function takes a vector of strings containing probability vectors and parses them, returning a matrix of the values, with column names given by the names of the states.

The function `parseProbVecRow()` is an internal function which parses a single row (after it has been split on the commas).

Value

A matrix containing the values. The rows correspond to the elements of pVec. The columns correspond to the state names.

Author(s)

Russell Almond

References

<http://research.ets.org/~ralmond/StatShop/dataFormats.html>

See Also

[readHistory](#)

Examples

```
parseProbVec(c(Good = "[High:.8,Med:.15,Low:.05]",
               Bad = "[High:.15,Med:.35,Low:.5]",
               Ugly = "[High:.01,Med:.09,Low:.9]"))
```

partialCredit

A link function based on the generalized partial credit model

Description

This function converts a matrix of effective theta values into a conditional probability table by applying the generalized partial credit model to each row of the table.

Usage

```
partialCredit(et, linkScale = NULL, obsLevels = NULL)
```

Arguments

et	A matrix of effective theta values. There should be one row in this table for each configuration of the parent variables of the conditional probability table and one column for each state of the child variables except for the last.
linkScale	Unused. For compatibility with other link functions.
obsLevels	A character vector giving the names of the child variable states. If supplied, it should have length <code>ncol(et)+1</code> .

Details

This function takes care of the third step in the algorithm of [calcDPCTable](#). Its input is a matrix of effective theta values (comparable to the last column of the output of [eThetaFrame](#)), one column for each of the child variable states (`obsLevels`) except for the last one. Each row represents a different configuration of the parent variables. The output is the conditional probability table.

Let X be the child variable of the distribution, and assume that it can take on M possible states labeled x_1 through x_M in increasing order. The generalized partial credit model defines a set of functions $Z_m(\theta_k)$ for $m = 2, \dots, M$, where

$$Pr(X \geq x_m | X \geq x_{m-1}, \theta_k) = \text{logit}^{-1} - D * Z_m(\theta_k)$$

The conditional probabilities for each child state is calculated by taking the differences between the curves.

The $K \times M - 1$ matrix `et` is the values of $Z_m(\theta_k)$. This function then performs the rest of the generalized partial credit model. The original Samejima (1969) development assumed that all of the functions $Z_m(\cdot)$ had the same linear form $a(\theta_k - b_m)$, with the b_m strictly increasing (note that in `CPTtools`, the states are ordered from highest to lowest, so that they should be strictly decreasing). This meant that the curves would never cross. The general notation of [calcDPCTable](#) does not ensure the curves do not cross, which could result in negative probabilities. This function handles this case by forcing negative probabilities to zero (and adjusting the probabilities for the other state to be properly normalized).

If supplied `obsLevels` is used for the column names.

Value

A matrix with one more column than `et` giving the conditional probabilities for each configuration of the parent variables (which correspond to the rows).

Note

The development here follows Muraki (1992) rather than Samejima (1969).

The `linkScale` parameter is unused. It is for compatibility with other link function choices.

Author(s)

Russell Almond

References

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (Draft) *Bayesian Networks in Educational Assessment*. Springer. Chapter 8.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, **16**, 159-176. DOI: 10.1177/014662169201600206

I also have planned a manuscript that describes these functions in more detail.

See Also

[eThetaFrame](#), [calcDPCTable](#), [partialCredit](#),

Examples

```
## Set up variables
skill1l <- c("High", "Medium", "Low")
correctL <- c("Correct", "Incorrect")
pcreditL <- c("Full", "Partial", "None")
gradeL <- c("A", "B", "C", "D", "E")

## Get some effective theta values.
et <- effectiveThetas(3)

partialCredit(matrix(et, ncol=1), NULL, correctL)

partialCredit(outer(et, c(Full=1, Partial=-1)), NULL, pcreditL)

partialCredit(outer(et, c(A=2, B=1, C=0, D=-1)), NULL, gradeL)
```

proflevelci

*Produce cumulative sum credibility intervals***Description**

Produces credibility intervals for hanging barplots. Assumes that each column represents a sum of proportions and produces corresponding intervals for the cumulative sums. Values hanging below and above the reference line are treated separately, and returned values below the reference are negative.

Usage

```
proflevelci(data, profindex, limits=list(lower=.025, upper=.975), a=.5, b=.5)
```

Arguments

data	A matrix of data values where each column refers to a bar in barplot. The values should be scaled so that the column sum is the number of individuals in that group.
profindex	The level in the chart which corresponds to the reference (proficiency) line. This should be a positive integer less than <code>nrow(data)</code> .
limits	The upper and lower credibility limits.
a	Value for the shape1 parameter of the beta prior.
b	Value for the shape2 parameter of the beta prior.

Details

For a stacked bar plot, the natural comparisons involve not category probabilities but the sum of the category probabilities up through the current bar. For hanging bar plots, this should go in both directions. So for example, if the categories are “Below Basic”, “Basic”, “Proficient”, and “Advanced”, and the zero line is to be put between “Basic” and “Proficient”, then we need credibility intervals for $\Pr(\text{“Basic” or “Below Basic”})$, $\Pr(\text{“Basic”})$, $\Pr(\text{“Proficient”})$, $\Pr(\text{“Proficient” or “Advanced”})$.

The `proflevelci` function splits the states up into those above the line and below the line using `profindex`. It then generates credibility intervals using [betaci](#) for the cumulative sums in each group. The primary purpose is to create confidence intervals for stacked bar charts (see [compareBars2](#)).

Value

A list of data sets of the same length as the `limits` argument. Each data set has the same shape as the `data` argument and represents a quantile of the data associated with the value in `limits`. With the default limits of `lower` and `upper`, the result is a list of two elements

<code>lower</code>	Gives the lower bounds of the confidence interval.
<code>upper</code>	Gives the upper bounds of the confidence interval.

Author(s)

Russell Almond

See Also

[betaci](#), [compareBars2](#)

Examples

```

margins <- data.frame (
  Trouble=c(Novice=19,Semester1=24,Semester2=28,Semester3=20,Semester4=9),
  NDK=c(Novice=1,Semester1=9,Semester2=35,Semester3=41,Semester4=14),
  Model=c(Novice=19,Semester1=28,Semester2=31,Semester3=18,Semester4=4)
)

proflevelci(margins,3,limits=c(lower=.025,upper=.975))

```

readHistory	<i>Reads a file of histories of marginal distributions.</i>
-------------	---

Description

In running a typical Bayes net engine, as each piece of evidence comes in, updated marginal distributions for several variables are output. This function reads a such a log, expressed as a comma separated value file, and creates a data structure suitable for doing weight of evidence analyses.

Usage

```
readHistory(csvfile)
```

Arguments

csvfile	A name of a CSV file containing StatShop marginal distributions, one per row. Expects a column named "Item" from which the names are taken and "Result" which contains the values.
---------	--

Details

This works with an excerpted log from a StatShop/ACED interaction. In this case the *Item* column should contain the name of the item presented at this iteration, possibly with an .xml suffix. The *Results* column should contain a probability vector of the form: [High:0.527,Medium:0.447,Low:0.025]. This function parses the CSV file and creates a matrix with rows corresponding to the rows in the CSV file and values from the probability vectors.

Value

A matrix whose column names are taken from the probability vectors and row names are taken from the *Item* field with the .xml suffix removed.

Author(s)

Russell Almond

References

<http://research.ets.org/~ralmond/StatShop/dataFormats.html>

See Also

[parseProbVec](#), [woeHist](#)

Examples

```
## Not run:
  allcorrect <- parseProbVec("CorrectSequence.csv")

## End(Not run)
```

rescaleTable	<i>Rescales the numeric part of the table</i>
--------------	---

Description

Takes a table representing a conditional probability distribution or a set of hyper-Dirichlet parameters and rescales the numeric part of the table. The function `rescaleTable()` scales the table by `scaleFactor`, the function `normalizeTable()` scales the function by the sum of the rows, making the result a conditional probability table.

Usage

```
rescaleTable(table, scaleFactor)
normalizeTable(table)
```

Arguments

<code>table</code>	A data frame describing a conditional probability table. Assumes that the conditions are expressed as factor variables, and all numeric columns represent states of the child variable.
<code>scaleFactor</code>	A scalar or vector of length equal to the number of rows of <code>table</code> .

Details

For `rescaleTable()`, every numeric column of `table` is multiplied by `scaleFactor`. This can be used to create a set of hyper-Dirichlet parameters by multiplying a conditional probability table by the effective sample size.

For `normalizeTable()`, the `scaleFactor` is set to be $1/\text{rowSums}(\text{table})$ (excluding the factor variables) so that the resulting table is a proper conditional probability table.

Value

A data frame of the same shape as `table` with the numeric entries suitably scaled.

Note

The function `scaleTable` does a similar rescaling, only it works with a separate ‘Sum’ and ‘Scale’ columns in the table.

Author(s)

Russell Almond

See Also

[getTableStates](#), [scaleTable](#)

Examples

```
#conditional table
X2.ptf <- data.frame(Theta=c("Expert","Novice"),
                    correct=c(4,2),
                    incorrect=c(2,4))

X2.t99 <- rescaleTable(X2.ptf,99/6) #Reweight to effective samples size of 99
X2.t31 <- rescaleTable(X2.ptf,c(3,1)) #Weight expert prior 3 times more than
                                     #novice prior.
X2.dtf <- normalizeTable(X2.ptf)

#Unconditional table
Theta.ptf <- data.frame(Expert=3,Novice=3)
Theta.t100 <- rescaleTable(Theta.ptf,100/6) #Reweight to effective
                                             #sample size of 100
Theta.dtf <- normalizeTable(Theta.ptf)
```

scaleMatrix

Scales a matrix to have a unit diagonal

Description

Creates a correlation matrix from a covariance matrix by scaling rows and columns to have a unit diagonal. Also can be used to create a partial correlation matrix from an inverse covariance/correlation matrix.

Usage

```
scaleMatrix(X)
```

Arguments

X A square, positive definite matrix (covariance matrix).

Details

Divides rows and columns by square root of the diagonal elements.

Value

A matrix of the same size and shape as the original with a unit diagonal.

Author(s)

Russell Almond

Examples

```
data(MathGrades)

## Create a correlation matrix from a covariance matrix.
round(scaleMatrix(MathGrades$var),2) == MathGrades$cor

## Create a partial correlation matrix from a correlation matrix
round(scaleMatrix(solve(MathGrades$cor)),2) == MathGrades$pcor
##Note: Some of these tests may return false due to machine precision
##issues.
```

scaleTable	<i>Scales a table according to the Sum and Scale column.</i>
------------	--

Description

Takes a matrix or vector with a Sum and Scale column and rescales it by multiplying each remaining element by the value of $Scale/Sum$ for that row.

If the last two rows are not named Sum and Scale then it simply returns its argument.

Usage

```
scaleTable(table)
```

Arguments

table	A matrix or vector in which the last two columns are named "Scale" and "Sum".
-------	---

Details

The parameters of a Dirichlet distribution can be stored in two ways, one is to have each cell in the table represent a pseudo count. The other was is to have each row represent a probability vector and use an additional pseudo sample size (the Scale column). If the probability vector is reported in a some other metric (say as a percentage or as a fraction of some smaller sample) the the Sum column is used to store the row sum.

Value

Rescaled table with Sum and Scale columns removed. This makes some attempt to preserve the type of the table argument as a matrix, row vector or numeric object.

Note

Used by the function compareDS to compare tables which may be in different formats.

Author(s)

Russell Almond

References

<http://research.ets.org/~ralmond/StatShop/dataFormats.html>

Examples

```
c1 <- matrix(c(70,20,10,10,20,70),nrow=2,byrow=TRUE,
             dimnames=list(NULL,c("H","M","L")))
s1 <- matrix(c(7,2,1,10,100,1,2,7,10,100),nrow=2,byrow=TRUE,
             dimnames=list(NULL,c("H","M","L","Sum","Scale")))

## 1 row matrixes need special handling (c1[1,] is a vector not a matrix)
c1r1 <- matrix(c1[1,],nrow=1,dimnames=list(NULL,c("H","M","L")))
s1r1 <- matrix(s1[1,],nrow=1,dimnames=list(NULL,c("H","M","L","Sum","Scale")))

stopifnot(
  identical(c1,scaleTable(s1)),
  identical(c1[1,],scaleTable(s1[1,])),
  identical(c1r1,scaleTable(s1r1))
)

# This should have no effect when run on matrixes without the Sum and
# Scale column.
stopifnot(
  identical(c1,scaleTable(c1)),
  identical(c1[1,],scaleTable(c1[1,])),
  identical(c1r1,scaleTable(c1r1))
)
```

stackedBarplot

Produces a hanging barplot

Description

This produces a series of stacked bar plots staggered so that the baseline corresponds to a particular state level. This is primarily designed for producing plots of probability vectors coming out of Bayes net scoring.

Usage

```
stackedBarplot(height, width = 1, space = 0.2, offset = 0, names.arg = NULL,
  legend.text = NULL, horiz = FALSE, density = NULL,
  angle = 45, col = NULL, border = par("fg"),
  main = NULL, sub = NULL, xlab = NULL, ylab = NULL,
  xlim = NULL, ylim = NULL, xpd = TRUE, axis = TRUE,
  axisnames = TRUE, cex.axis = par("cex.axis"),
  cex.names = par("cex.axis"), newplot = TRUE,
  axis.lty = 0, ...)
```

Arguments

height	A matrix giving the heights of the bars. The columns represent bars, and the rows represent groups within each bar.
width	A numeric vector of length equal to the number of columns in height giving the width of the bars. If value is shorter than number of columns in height it is recycled to produce the correct length.
space	A numeric vector of length equal to the number of columns in height giving the space between the bars. If value is shorter than number of columns in height it is recycled to produce the correct length.
offset	A numeric vector of length equal to the number of columns in height giving distance by which the bars should be offset from the zero line on the axis. Setting this to a non-zero value produces a hanging barplot.
names.arg	If not missing, used as axis labels (see axis).
legend.text	If not null, a legend is generated (see legend).
horiz	A logical value. If true, stacked bars are printed horizontally instead of vertically.
density	Density of shading lines (see rect). This should be a scalar or a vector of length equal to the number of rows of height.
angle	Angle of shading lines (see rect). This should be a scalar or a vector of length equal to the number of rows of height.
col	Color used for each bar. This should be a scalar or a vector of length equal to the number of rows of height. If not supplied a grayscale gradient is built.
border	Color for the rectangle borders (see rect). This should be a scalar or a vector of length equal to the number of rows of height.
main	Main title for plot (see title).
sub	Subtitle for plot (see title).
xlab	X-axis label for plot (see title).
ylab	Y-axis label for plot (see title).
xlim	Limits in user co-ordinates for the x-axis. Should be a vector of length 2.
ylim	Limits in user co-ordinates for the y-axis. Should be a vector of length 2.
xpd	A logical value controlling clipping. (see par).
axis	A logical value. If true, a numeric scale is printed on the appropriate axis.

<code>axisnames</code>	A logical value. If true, column names are printed on the appropriate axis.
<code>cex.axis</code>	Character size used for the numeric axis labels (see axis).
<code>cex.names</code>	Character size used for the text (column names) axis labels (see axis).
<code>newplot</code>	A logical value. If true a new graphics region is created. If false, the plot is placed on top of the existing graphics region.
<code>axis.lty</code>	A value passed as the <code>lty</code> argument to axis when plotting the text (column name) axis.
<code>...</code>	Other graphics parameters passed to rect , axis and title .

Details

This is a more detailed version of the [stackedBars](#) graph which allows finer control. It is used mainly by [compareBars](#).

There are some differences from [stackedBars](#). First, height can be any value, not just a vector of probability. Second, offset is given as a numeric value in the units of height, rather than as an index into the array of heights. Most of the rest of the arguments merely expose the graphical arguments to the user.

Value

The midpoints of the bars are returned invisibly.

Author(s)

Russell Almond

See Also

[compareBars](#), [colspread](#), [buildMarginTab](#), [marginTab](#), [barplot](#), [stackedBars](#)

Examples

```

margins <- data.frame (
  Trouble=c(Novice=.19,Semester1=.24,Semester2=.28,Semseter3=.20,Semester4=.09),
  NDK=c(Novice=.01,Semester1=.09,Semester2=.35,Semseter3=.41,Semester4=.14),
  Model=c(Novice=.19,Semester1=.28,Semester2=.31,Semseter3=.18,Semester4=.04)
)
margins <- as.matrix(margins)
baseline <- apply(margins[1:2,],2,sum)

stackedBarplot(margins,offset=-baseline,
  main="Marginal Distributions for NetPASS skills",
  sub="Baseline at 2nd Semester level.",
  col=hsv(223/360,.2,0.10*(5:1)+.5))

```

stackedBars

Produces a stacked, staggered barplot

Description

This produces a series of stacked bar plots staggered so that the baseline corresponds to a particular state level. This is primarily designed for producing plots of probability vectors coming out of Bayes net scoring.

Usage

```
stackedBars(data, profindex, ...,
            ylim = c(min(offsets) - 0.25, max(1 + offsets)),
            cex.names = par("cex.axis"), percent=TRUE,
            digits = 2*(1-percent), labrot=FALSE)
```

Arguments

data	A data.frame where each column is a probability vector.
profindex	The index of the proficiency which should be used as a baseline.
...	Graphical arguments passed to barplot.
ylim	Default limits for Y axis.
cex.names	Magnification for names.
percent	Logical value. If true data values are treated as percentages instead of probabilities.
digits	Number of digits for overlaid numeric variables.
labrot	If true, labels are rotated 90 degrees.

Details

This plot type assumes that each column in its first argument is a probability vector. It then produces a stacked bar for each column. The baseline of the bar is offset by the probability for being in the category marked by profindex or below.

The probability values are overlaid on the bars.

Author(s)

Russell Almond

References

This plot type was initially developed in Jody Underwood's Evolve project.

Almond, R. G., Shute, V. J., Underwood, J. S., and Zapata-Rivera, J.-D (2009). Bayesian Networks: A Teacher's View. *International Journal of Approximate Reasoning*. **50**, 450-460.

See Also

[compareBars](#), [colspread](#), [buildMarginTab](#), [marginTab](#), [barplot](#), [stackedBarplot](#)

Examples

```

margins <- data.frame (
  Trouble=c(Novice=.19,Semester1=.24,Semester2=.28,Semester3=.20,Semester4=.09),
  NDK=c(Novice=.01,Semester1=.09,Semester2=.35,Semester3=.41,Semester4=.14),
  Model=c(Novice=.19,Semester1=.28,Semester2=.31,Semester3=.18,Semester4=.04)
)

stackedBars(margins,3,
  main="Marginal Distributions for NetPASS skills",
  sub="Baseline at 3rd Semester level.",
  cex.names=.75, col=hsv(223/360,.2,0.10*(5:1)+.5))

stackedBars(margins,3,
  main="Marginal Distributions for NetPASS skills",
  sub="Baseline at 3rd Semester level.",
  percent=FALSE,digits=2,
  cex.names=.75, col=hsv(223/360,.2,0.10*(5:1)+.5))

```

structMatrix

Finds graphical structure from a covariance matrix

Description

This function finds an undirected graphical representation of a multivariate normal distribution with the given covariance matrix, by associating edges with non-zero entries. Graphical structure is given as an adjacency matrix.

Usage

```
structMatrix(X, threshold = 0.1)
```

Arguments

X	A variance matrix.
threshold	A numeric value giving the threshold for a value to be considered “non-zero”.

Details

For a multivariate normal model, zero entries in the inverse covariance matrix correspond to conditional independence statements true in the multivariate normal distribution (Whitaker, 1990; Dempster, 1972). Thus, every non-zero entry in the inverse correlation matrix corresponds to an edge in an undirected graphical model for the structure.

The `threshold` parameter is used to determine how close to zero a value must be to be considered zero. This allows for both estimation error and numerical precision when inverting the covariance matrix.

Value

An adjacency matrix of the same size and shape as `X`. In this matrix `result[i, j]` is TRUE if and only if Node *i* and Node *j* are neighbors in the graph.

Note

Models of this kind are known as “Covariance Selection Models” and were first studied by Dempster (1972).

Author(s)

Russell Almond

References

Dempster, A.P. (1972) Covariance Selection. *Biometrics*, **28**, 157–175.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

See Also

[scaleMatrix](#), [mcSearch](#), [buildParentList](#)

Examples

```
data(MathGrades)

MG.struct <- structMatrix(MathGrades$var)
```

woeBal

Weight of Evidence Balance Sheet

Description

Creates a weight of evidence balance sheet from a history of marginal distributions.

Usage

```
woeBal(hist, pos, neg, obs=NULL, title = "Evidence Balance Sheet",
       col = rev(colorsread("slategray", ncol(hist), maxsat=TRUE)),
       posCol="cyan", negCol="red", stripCol=c("white", "lightgray"),
       lcex = 0.65)
```


Arguments

hist	A matrix whose rows represent time points (after tests) and columns represent probabilities.
pos	Names or numbers of states which should be regarded as “positive”
neg	Names or numbers of states which should be regarded as “negative”
obs	An optional character vector of the same length as the number of rows of hist giving the value observed at each step
title	Title for plot
col	A list of color values for probability bars.
posCol	The color to be used for bars showing positive weights of evidence.
negCol	The color to be used for bars showing negative weights of evidence.
stripCol	The colors to be used for the time step labels. Setting this to a vector of two colors creates alternate color stripes. Set this to "white" to disable that effect.
lcex	Character expansion size for labels.

Details

This constructs a weight of evidence balance sheet (Madigan, Mosurski, and Almond, 1997) showing the changes to the probability distribution and weight of evidence for each change in the probability. The probabilities are given in the hist argument in which each row should be a probability distribution for the target variable. The labels for the plot are taken from the row labels of the hist argument.

Madigan, Mosurski and Almond (1997) note that the definition of weight of evidence is somewhat problematic if the hypothesis variable is not binary. In that case, they recommend partitioning the states into a *positive* and *negative* set. The pos and neg are meant to describe that partition. They can be any expression suitable for selecting columns from the hist matrix. This function calls woeHist() to calculate weights of evidence.

The row names of hist are printed left-justified in the leftmost column. If observed values (obs) are supplied, they are printed right justified in the same column.

Value

The midpoints of the bars (see [barplot](#)) are returned invisibly.

Side Effects

Starts a new plotting page and creates three side-by-side plots, one for the labels, one for the probability bars and one for the weight of evidence bars.

Author(s)

Russell Almond

References

Good, I. (1971) The probabilistic explication of information, evidence, surprise, causality, explanation and utility. In *Proceedings of a Symposium on the Foundations of Statistical Inference*. Holt, Rinehart and Winston, 108-141.

Madigan, D., Mosurski, K. and Almond, R. (1997) Graphical explanation in belief networks. *Journal of Computational Graphics and Statistics*, **6**, 160-181.

Almond, R. G., Kim, Y. J., Shute, V. J. and Ventura, M. (2013). Debugging the Evidence Chain. In Almond, R. G. and Mengshoel, O. (Eds.) *Proceedings of the 2013 UAI Application Workshops: Big Data meet Complex Models and Models for Spatial, Temporal and Network Data (UAI2013AW)*, 1-10. <http://ceur-ws.org/Vol-1024/paper-01.pdf>

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. and Yan, D. (2015) *Bayesian Networks in Educational Assessment*. Springer. Chapter 7.

See Also

[readHistory](#), [woeHist](#), [barplot](#), [Colors](#)

Examples

```
sampleSequence <- read.csv(paste(library(help="CPTtools")$path,
                                "testFiles", "SampleStudent.csv",
                                sep=.Platform$file.sep),
                           header=TRUE, row.names=1)

woeBal(sampleSequence[,c("H", "M", "L")], c("H"), c("M", "L"), lcex=1.25)
woeBal(sampleSequence[,c("H", "M", "L")], c("H"), c("M", "L"),
        obs=sampleSequence[, "Acc"], lcex=1.25)
```

woeHist

Creates weights of evidence from a history matrix.

Description

Takes a matrix providing the probability distribution for the target variable at several time points and returns a weight of evidence for all time points except the first.

Usage

```
woeHist(hist, pos, neg)
```

Arguments

hist	A matrix whose rows represent time points (after tests) and columns represent probabilities.
pos	Names or numbers of states which should be regarded as “positive”
neg	Names or numbers of states which should be regarded as “negative”

Details

Good (1971) defines the *Weight Of Evidence (WOE)* as:

$$100 \log_1 0 \frac{\Pr(E|H)}{\Pr(E|\overline{H})} = 100 \left[\log_1 0 \frac{\Pr(H|E)}{\Pr(\overline{H}|E)} - \log_1 0 \frac{\Pr(H)}{\Pr(\overline{H})} \right]$$

Where \overline{H} is used to indicate the negation of the hypothesis. Good recommends taking the log base 10 and multiplying by 100, and calls the resulting units *centibans*. The second definition of weight of evidence as a difference in log odd leads naturally to the idea of an incremental weight of evidence for each new observation.

Following Madigan, Mosurski and Almond (1997), all that is needed to calculate the WOE is the marginal distribution for the hypothesis variable at each time point. They also note that the definition is somewhat problematic if the hypothesis variable is not binary. In that case, they recommend partitioning the states into a *positive* and *negative* set. The pos and neg are meant to describe that partition. They can be any expression suitable for selecting columns from the hist matrix.

Value

A vector of weights of evidence of length one less than the number of rows of hist (i.e., the result of applying `diff()` to the vector of log odds.)

Author(s)

Russell Almond

References

Good, I. (1971) The probabilistic explication of information, evidence, surprise, causality, explanation and utility. In *Proceedings of a Symposium on the Foundations of Statistical Inference*. Holt, Rinehart and Winston, 108-141.

Madigan, D., Mosurski, K. and Almond, R. (1997) Graphical explanation in belief networks. *Journal of Computational Graphics and Statistics*, **6**, 160-181.

See Also

[readHistory](#), [woeBal](#), [diff](#)

Examples

```
## Not run:  
  allcorrect <- parseProbVec("CorrectSequence.csv")  
  woeHist(allcorrect,c("High"),c("Medium","Low"))  
  woeHist(allcorrect,1:2,3)  
  
## End(Not run)
```

Index

- *Topic **category**
 - scaleTable, 74
 - *Topic **datasets**
 - ACED.scores, 8
 - MathGrades, 56
 - *Topic **distribution**
 - calcDDTable, 21
 - calcDNTable, 23
 - calcDPCTable, 26
 - calcDSllike, 30
 - calcDSTable, 32
 - calcNoisyAndTable, 35
 - calcNoisyOrTable, 37
 - Compensatory, 43
 - effectiveThetas, 46
 - eThetaFrame, 47
 - gradedResponse, 50
 - mapDPC, 54
 - OffsetConjunctive, 64
 - partialCredit, 67
 - profllevelci, 69
 - *Topic **graphics**
 - colorspread, 40
 - OCP, 62
 - *Topic **hplot**
 - compareBars, 41
 - stackedBarplot, 75
 - stackedBars, 78
 - woeBal, 80
 - *Topic **htest**
 - localDepTest, 52
 - *Topic **interface**
 - parseProbVec, 66
 - readHistory, 71
 - *Topic **manip**
 - areaProbs, 12
 - buildFactorTab, 15
 - buildParentList, 17
 - buildRegressions, 18
 - buildRegressionTables, 19
 - dataTable, 45
 - getTableStates, 49
 - mcSearch, 58
 - rescaleTable, 72
 - scaleMatrix, 73
 - structMatrix, 79
 - *Topic **math**
 - woeHist, 82
 - *Topic **multivariate**
 - mutualInformation, 59
 - *Topic **package**
 - CPTtools-package, 2
 - *Topic **tests**
 - betaci, 14
 - *Topic **utilities**
 - numericPart, 61
-
- ACED, 5
 - ACED (ACED.scores), 8
 - ACED.scores, 8
 - areaProbs, 4, 12
 - axis, 76, 77
 - barplot, 43, 77, 79, 81, 82
 - betaci, 5, 14, 62–64, 70
 - build2FactorTab (buildFactorTab), 15
 - buildFactorTab, 5, 15, 43, 53
 - buildMarginTab, 77, 79
 - buildMarginTab (buildFactorTab), 15
 - buildParentList, 4, 17, 19, 58, 59, 80
 - buildRegressions, 4, 18, 21
 - buildRegressionTables, 4, 19, 19
 - calcDDFrame (calcDDTable), 21
 - calcDDTable, 4, 21
 - calcDNFrame (calcDNTable), 23
 - calcDNllike, 25
 - calcDNllike (calcDSllike), 30

- calcDNTable, [4](#), [23](#), [23](#), [29](#), [31](#), [34](#), [37](#), [39](#), [45](#), [49](#), [66](#)
- calcDPCTable (calcDPCTable), [26](#)
- calcDPCTable, [3](#), [25](#), [26](#), [34](#), [37](#), [39](#), [45](#), [49](#), [51](#), [54–56](#), [66](#), [68](#)
- calcDSFrame (calcDSTable), [32](#)
- calcDSlike, [4](#), [30](#), [34](#), [46](#)
- calcDSTable, [4](#), [22](#), [23](#), [25](#), [27](#), [29](#), [31](#), [32](#), [37](#), [39](#), [45](#), [49](#), [65](#), [66](#)
- calcNoisyAndFrame (calcNoisyAndTable), [35](#)
- calcNoisyAndTable, [4](#), [35](#)
- calcNoisyOrFrame (calcNoisyOrTable), [37](#)
- calcNoisyOrTable, [4](#), [37](#), [37](#), [39](#)
- ciTest (localDepTest), [52](#)
- col2rgb, [40](#)
- colors, [41](#)
- colspread, [5](#), [40](#), [43](#), [77](#), [79](#)
- compareBars, [5](#), [16](#), [41](#), [41](#), [77](#), [79](#)
- compareBars2, [70](#)
- compareBars2 (compareBars), [41](#)
- Compensatory, [3](#), [22–25](#), [27](#), [29](#), [31](#), [33](#), [34](#), [43](#), [48](#), [49](#), [56](#), [66](#)
- Conjunctive, [22](#), [24](#), [27](#), [33](#), [48](#)
- Conjunctive (Compensatory), [43](#)
- CPTtools (CPTtools-package), [2](#)
- CPTtools-package, [2](#)
- data.frame, [61](#)
- data.matrix, [61](#)
- dataTable, [4](#), [31](#), [45](#)
- diff, [83](#)
- Disjunctive, [22](#), [24](#), [27](#), [33](#), [48](#)
- Disjunctive (Compensatory), [43](#)
- do.call, [28](#)
- effectiveThetas, [3](#), [4](#), [13](#), [22–25](#), [27](#), [29](#), [33](#), [34](#), [45](#), [46](#), [48](#), [49](#), [55](#), [66](#)
- eThetaFrame, [4](#), [25](#), [29](#), [31](#), [34](#), [45](#), [47](#), [51](#), [66](#), [68](#)
- expand.grid, [23](#), [25](#), [27](#), [29](#), [34](#), [37](#), [39](#), [49](#)
- factorPart (numericPart), [61](#)
- getTableParents (getTableStates), [49](#)
- getTableStates, [4](#), [49](#), [73](#)
- gradedResponse, [3](#), [28](#), [29](#), [50](#), [51](#), [56](#)
- legend, [42](#), [76](#)
- localDepTest, [5](#), [52](#)
- mantelhaen.test, [53](#)
- mapDPC, [3](#), [54](#)
- marginTab, [77](#), [79](#)
- marginTab (buildFactorTab), [15](#)
- MathGrades, [5](#), [56](#)
- matrix, [61](#)
- mcSearch, [4](#), [17](#), [18](#), [58](#), [80](#)
- MutualInfo, [5](#)
- mutualInformation, [5](#), [59](#)
- normalizeTable (rescaleTable), [72](#)
- numericPart, [4](#), [61](#)
- OCP, [5](#), [15](#), [62](#)
- OCP2 (OCP), [62](#)
- OffsetConjunctive, [3](#), [24](#), [25](#), [28](#), [29](#), [31](#), [33](#), [34](#), [45](#), [48](#), [49](#), [56](#), [64](#)
- OffsetDisjunctive, [24](#), [28](#), [33](#), [48](#)
- OffsetDisjunctive (OffsetConjunctive), [64](#)
- optim, [55](#), [56](#)
- par, [42](#), [62](#), [63](#), [76](#)
- parseProbVec, [66](#), [71](#)
- parseProbVecRow (parseProbVec), [66](#)
- partialCredit, [3](#), [28](#), [29](#), [56](#), [67](#), [68](#)
- Peanut, [3](#), [6](#)
- plot.window, [62](#), [63](#)
- proflevelci, [69](#)
- pvecToCutpoints, [20](#)
- pvecToCutpoints (areaProbs), [12](#)
- pvecToMidpoints (areaProbs), [12](#)
- readHistory, [67](#), [71](#), [82](#), [83](#)
- rect, [76](#), [77](#)
- rescaleTable, [4](#), [50](#), [72](#)
- rgb2hsv, [40](#)
- RNetica, [6](#)
- scaleMatrix, [4](#), [73](#), [80](#)
- scaleTable, [4](#), [72](#), [73](#), [74](#)
- stackedBarplot, [75](#), [79](#)
- stackedBars, [5](#), [16](#), [43](#), [77](#), [78](#)
- structMatrix, [4](#), [18](#), [59](#), [79](#)
- table, [45](#), [46](#), [52](#), [60](#)
- title, [42](#), [62](#), [63](#), [76](#), [77](#)

UCBAdmissions, [53](#)

woeBal, [5](#), [80](#), [83](#)

woeHist, [5](#), [71](#), [82](#), [82](#)